Torture and the Commitment Problem[*]

Sandeep Baliga [†] Jeffrey C. Ely [‡]

November 2, 2015

Abstract

We study torture as a mechanism for extracting information from a suspect who may or may not be informed. We show that a standard rationale for torture generates two commitment problems. First, the principal would benefit from a commitment to torture a suspect he knows to be innocent. Second, the principal would benefit from a commitment to limit the amount of torture faced by the guilty. We analyze a dynamic model of torture in which the credibility of these threats and promises is endogenous. We show that these commitment problems dramatically reduce the value of torture and can even render it completely ineffective. We use our model to address questions such as the effect of enhanced interrogation techniques, rights against indefinite detention, and delegation of torture to specialists.

Keywords: commitment, waterboarding, sleep deprivation, ratchet effect.

[*] We thank many seminar audiences and especially Nageeb Ali, Simon Board, Laura Doval, Navin Kartik, Roger Lagunoff and Pierre Yared for useful comments. We also thank the Editor, Francesco Caselli, and four referees for helpful suggestions.

[†] Kellogg Graduate School of Management, Northwestern University. baliga@kellogg.northwestern.edu

[‡] Department of Economics, Northwestern University. jeffely@northwestern.edu.

1 Introduction

A major terrorist attack is planned for some time in the future. A suspect with potential intelligence about the impending attack awaits interrogation. Perhaps the suspect was caught in the wrong place at the wrong time and is completely innocent. He may even be a terrorist but have no useful information about the imminent attack. But there is another possibility: the suspect is a senior member of a terrorist organization and was involved in planning the attack. If the suspect yields actionable intelligence, the terrorist attack can be averted or its impact reduced. In this situation, suppose torture is the only instrument available to obtain information.

Uncertainty about how much useful intelligence a prisoner possesses is commonplace,[1] and there is a lively debate about whether torture should be used to extract information. There is a dilemma: the suspect's information may be valuable but torture is costly and abhorrent to society. Walzer (1973) famously argues that a moral decision maker facing this dilemma should use torture because the value of saving many lives outweighs the costs.[2] Dershowitz (2002) goes further and argues torture should be legalized.

If this rationale can be used to justify starting torture in the first place, it can also be used to justify continuing or ending torture once it has begun. Then, two commitment problems arise. First, if torture of a high value target is meant to stop after some time, there is an incentive to renege and continue in order to extract even more information. After all, innocent lives are at stake and if the threat of torture saves more of them, it is right to continue whatever promise was made.[3] Second, if after enough

[1] For example, in many interrogations in Iraq a key question is whether a detainee is a low level technical operative or a senior Al Qaeda leader (see Alexander and Bruning (2008)).

[2] "[C]onsider a politician who has seized upon a national crisis-a prolonged colonial war-to reach for power.....Immediately, the politician goes off to the colonial capital to open negotiations with the rebels. But the capital is in the grip of a terrorist campaign, and the first decision the new leader faces is this: he is asked to authorize the torture of a captured rebel leader who knows or probably knows the location of a number of bombs hidden in apartment buildings around the city, set to go off within the next twenty-four hours. He orders the man tortured, convinced that he must do so for the sake of the people who might otherwise die in the explosions..."

[3] For example, as commentator Liz Cheney asks, "Mr. President, in a ticking time-

resistance we learn that the suspect is likely a low value target, there is an incentive to stop. With limited personnel to carry out interrogation and verify elicited information, it is better to redeploy assets to interrogate another suspect who might be informed rather than continue with one likely to have no useful intelligence.[4] And since torture is abhorrent, to inflict it on an uninformed suspect cannot be justified. Both of these commitment problems encourage the informed suspect to resist torture. The first problem discourages early confession because the suspect anticipates that it would only lead to further torture. The second problem also discourages confession as silence may hasten the cessation of torture.

What is the value of torture to a principal when these two commitment problems are present? We study a dynamic model of torture where a suspect/agent faces a torturer/principal. The suspect may have information that is valuable to the principal – he might know where bombs are hidden or locations of various persons of interest. We study the value of torture as an instrument for extracting that information. This information extraction rationale is invoked to justify torture in contemporary policy debates and hence this is the scenario on which we focus. We emphasize that we are not studying torture as a means of terrorizing or extracting a false confession for its own sake. While it is clear that torture has been used throughout history for these means, and even as an end in itself, the purpose of our study is to focus on the purely instrumental value of torture.

Each period, the principal decides whether to demand some information from the suspect backed by the threat of torture. The suspect either reveals *verifiable* information or suffers torture. For example, an agent can offer a location of a target such as bomb or a wanted terrorist and the principal can check whether there is in fact a target at the reported address. An informed agent can always reveal a true location while an un-

bomb scenario, with American lives at stake, are you really unwilling to subject a terrorist to enhanced interrogation to get information that would prevent an attack?”(Leibovich (2009)) This argument for torture is even stronger if a subject is known to be an informed terrorist.

[4] In the report on the prison at Abu Ghraib, Major General George Fay (Fay, 2004, p37) reports,”Large quantities of detainees with little or no intelligence value swelled Abu Ghraib’s population and led to a variety of overcrowding difficulties. Already scarce interrogator and analyst resources were pulled from interrogation operations to identify and screen increasing numbers of personnel whose capture documentation was incomplete or missing.” Hence, supply of experienced personnel is a binding constraint on interrogation.

informed agent can at best give a false address. Torture inflicts costs on both the agent and the principal. These costs are proportional to the period length to which the principal can commit to torture if information is not forthcoming. We study two versions of this model. In the *perpetual threat scenario*, there is an infinite horizon and the terrorist event can occur with positive probability at any time. In the *ticking time-bomb scenario*, as in canonical justifications for the use of torture, there is a "ticking time-bomb": the principal wants to extract as much information as possible prior to a fixed terminal date when the attack will take place. The interrogation process continues until either all of the information is extracted or time runs out. We also study some extensions including the use of enhanced interrogation techniques.[5]

FIXME THIS PARAGRAPH

The key intuition for our argument is simple:

If the suspect reveals some information, the principal will continue to extract more information under the threat of torture. If the suspect stays silent, the principal cannot credibly commit to costly torture of a resistant suspect and will eventually stop. This gives the informed suspect the incentive to resist torture because, if he yields, he gives up even more and, if he does not, he escapes torture and retains his information. Finally, since the informed suspect is resolute, there is no incentive for the principal to torture. In fact, in the perpetual threat scenario, if players are sufficiently patient, there is an equilibrium where the suspect never confesses on the equilibrium path and the principal does not torture at all. This is obviously the worst equilibrium for the principal. Our main contribution is to show this logic is also reflected in the *best* equilibrium for the principal when the principal can continuously revisit his torture decision.

The principal's commitment problem means he cannot credibly commit to torture in any period where the expected marginal benefits - the chance of extra information - are outweighed by costs of torture. The "stick" of torture to incentivize the agent to yield is also a stick to the principal and so he must be rewarded by enough of a "carrot" of expected information. This implies that, even in the perpetual threat scenario, the principal must eventually stop torturing a suspect who has not confessed. At some point, either the principal faces a suspect who is likely to be uninformed or he faces a suspect whose equilibrium strategy is not to confess

[5] In the Conclusion, we also offer other applications of our model.

with high probability even if he is informed. In either case, the benefits of torturing are outweighed by the costs so the principal must stop torture. This means that a suspect who never concedes for long enough ultimately escapes torture. After a finite number of periods, the best equilibrium resembles the worst equilibrium.

Of course, if the total amount of time the principal tortures a resistant suspect is large, the threat of lengthy torture is he resists might still persuade the informed suspect to yield a significant amount of information. But the principal's commitment problem also undermines the total length of time torture can be credibly threatened even in the best equilibrium. The expected marginal benefits of torturing a resistant suspect in any period depend on the probability the suspect is informed and the probability he confesses in that period. These variables are connected across periods - for instance, if the probability of concession by an informed suspect is high one period, then the probability a resistant suspect is informed must be low in the next. We show this means that there cannot be many periods where the expected marginal benefits of torturing a resistant suspect are high. Hence, he cannot be credibly threatened with torture for many periods. In fact, we derive an upper bound on the number of periods that torture can be credibly threatened as a function of the principal's prior, the per period costs of torture to the principal and the per period costs of torture to the agent (the last determines the maximum amount of information an informed suspect will ever give up in a period and enters the expected benefits of torture). Even if the principal has more time available to torture, he will not use it. Hence, laws against indefinite detention do not lessen the value of torture.

The principal might be able to revisit his torture policy continuously. For example, in his account of enhanced interrogations in Iraq, Alexander and Bruning (2008) describes frequent decisions as to whether to continue with a suspect or switch to a new one. Hence, we study the best equilibrium for the principal as the period length becomes small while keeping flow costs and the principal's prior fixed. Then, a shorter period length has no impact if the principal can fully commit but with limited commitment there are more points in time for the principal to re-evaluate his torture decision. Therefore, we gain a deeper understanding of the commitment problem and our results shed light on the value of torture when decisions can be made almost continuously.

If the principal's equilibrium strategy is simply to torture for more periods in the same length of physical time, the value of torture does not change. Yet we show that the value of torture shrinks to zero when the period length shortens. In any time interval, the principal's marginal expected benefits from torture must be high at more and more points in time, otherwise he does not torture a suspect who has not confessed. But as we argued above, benefits can be high for only a few periods which now correspond to a shorter interval of physical time. Knowing that he will face the costs of torture for only a short time if he does not confess, an informed agent will only give up a small amount of information. Hence, as the period length goes to zero, so does the amount of information an informed suspect yields.

Many of these properties apply to both the perpetual threat scenario and the canonical ticking time-bomb scenario but there are differences. For example, it might seem that the known date at which the ticking time-bomb explodes may help the principal to commit and hence increase the value of torture. But we show the value of torture is even lower for the principal in the ticking time-bomb scenario. In the perpetual threat scenario, the worst equilibrium for the principal can be used to increase the expected costs of a deviation and purchase some commitment. But there is a unique equilibrium in the ticking time-bomb scenario and there is less to prevent the principal from deviating. This has two implications. First, the principal is forced to use the date at which the time-bomb explodes as a commitment device to stop torturing and torture begins only towards the end. If payoffs are discounted, this reduces the principal's payoff relative to the perpetual threat scenario. Also, since the principal cannot be punished for suspending torture, the expected marginal benefits from torture must be even higher in every period where torture to be credibly employed on a resistant suspect. But this reduces the number of periods that torture can be credibly threatened and hence the value of torture is even lower in the ticking time-bomb scenario.

To summarize: Even in the best circumstances, the principal can credibly threaten to torture a resistant suspect for only so many periods. This is because the principal is comparing the expected marginal benefits of torture with the expected marginal costs. The benefits depend on the probability the resistant suspect is informed and the probability he will confess in that period. These can be high only for a few periods. If the principal can revisit his torture policy continuously, torture cannot be credibly threatened on a resistant suspect and so the informed suspect will not give up anything information.

This conclusion stands in stark contrast to our first extension where assume torture can be outlawed and allow the principal to promise money in return for information but to commit to pay for only one period at a time. Giving the suspect a carrot in return for information is also aligned with the principal's incentives as he receives information in return for a costly transfer to the agent. This alignment is not compromised even if payments have to occur frequently. Therefore, we show that monetary payments are effective in extracting information when only small payments can be made near continuously. This shows that limited commitment does not undermine *all* incentive schemes and identifies the superiority of monetary transfers over costly interrogation as a tool of information extraction when there is limited commitment.

We consider two other extensions. First, the principal may need to know *all* the agent's information before it is useful. We capture this by allowing the principal's value of information to be convex in the quantity extracted. We show this can only reduce the value of torture to the principal. This is because the stakes of releasing the last bit of information dramatically reduces the payoff to the informed agent compared to the cost of torture. So, he would never release it and this undermines the principal's incentive to torture in the first place. Second, to evaluate the use of "enhanced interrogation techniques" we study a model in which the principal can choose either a mild torture technology ("sleep deprivation") or a harsher one ("waterboarding"). The mild technology extracts less information per period but is less costly so that in some cases the principal may prefer it over the harsh technology. We show how the existence of the enhanced interrogation technique compromises the use of the mild technology. Once the suspect starts talking under the threat of sleep deprivation, the principal cannot commit not to increase the threat and use waterboarding to extract more information. This reduces the suspect's incentive to concede in the first place lowering the principal's overall payoff in the ticking time-bomb scenario.

Finally, we discuss the difficulties with standard solutions to the commitment problem. For example, delegation can often solve commitment problems and we have identified two that limit the value of torture. Indeed, delegating torture to a specialist with a preference for torture ameliorates one commitment problem: he is willing to continue even if the probability the suspect is informed is zero. This means the informed suspect can concede information with probability one in equilibrium. On the other hand the specialist cannot commit to limit torture. Indeed, the specialist will torture the agent in all periods which are not utilized for information extraction. If the time horizon is long, the value of torture to the principal is lower with delegation than without. Moreover, there is a fundamental problem with using delegation to resolve commitment problems particularly in the torture environment: As torture is carried out in secret and is unverifiable, the principal cannot commit to keep the specialist employed. As soon as the agent does not yield information, the principal intervenes and stops torture. Then, the commitment problem reappears.

Strategic advice to suspects and principals resonates with the key properties of our analysis. An Al Qaeda manual describes torture techniques and how to fight them (Post (2005)):

> "The brother may think that by giving a little information he can avoid harm and torture. However, the opposite is true. The torture and harm would intensify to obtain additional information, and that cycle would repeat. Thus, the brother should be patient, resistant, silent, and prayerful to Allah, especially if the security apparatus knows little about him."

The credible revelation of information leads to yet more intense torture while the only chance of escape comes from dissembling. This resembles not only the implications of the two basic commitment problems we study but implies a ratchet effect if the suspect talks.

Our analysis predicts that once an agent reveals information, torture is not utilized on the equilibrium path unless the agent stops cooperating. This policy is recommended in a C.I.A. interrogation manual:[6]

> "Once a confession is obtained, the classic cautions apply. The pressures are lifted enough so that the subject can provide information as accurately as possible. In fact, the relief granted the subject at this time fits neatly into the "questioning" plan. He is told that the changed treatment is a reward for truthfulness and evidence *that friendly handling will continue as long as he cooperates* [our emphasis]."

[6] CIA Human Resource Exploitation Manual.

If the suspect never cooperates at all, the canonical procedure is described by Alexander and Bruning (2008) which is based on the experience of an interrogator in Iraq (see pages 188-189 or 218 for example). After every period of interrogation when no credible information has been extracted, the key decision is whether to “retain and extract” or “transfer” the suspect out of the facility. These qualitative features also fit the predictions from our model. Moreover, the principal’s strategy described in these sources dovetails with the suspect’s strategy recommended in the Al Qaeda training manual and vice versa.

An age-old and yet contemporary argument warns of false confessions as suspects attempt to escape torture.[7] In our model, false confessions are equivalent to revealing no useful information at all. The concern about false confessions does not undercut the case against torture if only uninformed suspects make up evidence while informed suspects reveal it truthfully. But we show that when the principal has a commitment problem, *both* informed and uninformed players will yield false information. Thus, our model helps to clarify the logic of the common argument against torture and shows that it hinges on limited commitment.

We assume torture is costly. This cost can arise from a number of channels. First, the classical argument sees a moral cost arising from the repugnance of torture.[8] Second, the fact that torture is considered morally reprehensible begets laws against torture. Professional interrogators even

[7] Fifteen hundred years ago the Roman jurist Ulpian warned that torture might not generate truthful evidence for this reason (*Corpus Juris Civilis*, Dig. 48.18.1.23.) Lawrence Wilkerson, the former chief of staff at the State Department, reveals that the evidence linking Saddam to Al Qaeda was extracted by waterboarding suspect al-Libi (Wilkerson (2009).) He adds, “Of course later we learned that al-Libi revealed these contacts only to get the torture to stop. There in fact were no such contacts.”

[8] For example, St. Augustine (Augustine, 1825, Book 19, Chapter 6), “Of the error of human judgments when the truth is hidden. What shall I say of torture applied to the accused himself? He is tortured to discover whether he is guilty, so that, though innocent, he suffers a severe punishment for crime that is still doubtful, not because it is proved that he committed it, but because it is not known that he did not commit it. And through this ignorance of the judge, the innocent man suffers... And the judge thinks it not contrary to divine law that innocent witnesses are tortured in cases dealing with the crimes of others... or that the accused are put to the torture and, though innocent, make false confessions regarding themselves, and are punished; or that, though they be not condemned to die, they often die during the torture.” Here St. Augustine identifies the asymmetry of information between the principal and the agent as well as the moral repugnance of torture.

if they face no moral qualms themselves may fear prosecution if they actually use illegal methods rather than just threaten to use them. The U.S. policy of extraordinary rendition which brought terrorist suspects to neutral countries for interrogation is evidence of these types of costs and the incentive to reduce them. Third, using an interrogation technology - the interrogator, the holding cell etc. - on one suspect is costly if it precludes its use on someone else. This appears to be a significant practical concern.[9]

Before turning to the formal model, we take the opportunity to discuss related results. Our result that the principal's commitment power vanishes as the period length becomes small is reminiscent of results like the Coase conjecture for durable goods bargaining but the logic is very different. For example, in our ticking time bomb model, there is no discounting and a fixed finite horizon. In this setting a durable goods monopolist could secure at least the static monopoly price regardless of the way time is discretized (see for example Horner and Samuelson (2009)). The key feature that sets torture apart is that the agent can never be induced to concede a lot of information in a short period of time because this would be more costly than the threat of torture itself. As the period length shortens, the principal tortures for the same *number* of periods but this represents a smaller and smaller interval of real time. The total threat over that vanishing length of time is itself vanishing and hence so is the total amount of information the agent chooses to reveal.[10]

In reputation models, it is possible to obtain a lower bound on a long-run player's equilibrium payoff (see Kreps and Wilson (1982) and Fudenberg and Levine (1992, 1989)). Our model is distinguished from standard reputation models in two important respects. First, in our model there are two long-run players. Thus our conclusions do not follow from arguments based on learning rates as in Fudenberg and Levine (1992) which are the basis of most of the reputation results in the literature. An important exception is Myerson (1991) who studies an infinite horizon alternating-offer bargaining with two long run players. One player may be a commitment

[9] (Alexander and Bruning, 2008, p. 43) report "[The supervisor i]s not going to keep him and Abu Ali around much longer. . . They are not giving us anything, and the [Special Forces] guys bring in new catches every night."

[10] Suppose that the two parties are bargaining over the *rental rate* of a durable good which will perish after some fixed terminal date. As the terminal date approaches and no agreement has yet to be reached, the total gains from trade shrinks.

type who accepts an offer if and only if it is greater than some fraction of the surplus.

Unlike the bargaining game in Myerson (1991), torture is a dynamic game where the state variable is the amount of information yet to be revealed. This is the second key difference with the standard reputation literature. To see its role note that in any reputational bargaining model with a finite deadline (as we have in the ticking time bomb scenario) the payoff of the uninformed player (the principal in our model) is bounded below by the payoff he would get by waiting until the deadline and making a final offer that will be accepted by the uncommitted type of opponent. When the probability of the uncommitted type is large, this lower bound is large and unaffected by the length of the period between offers. By contrast in our model, regardless of the type distribution, the principal's payoff shrinks to zero as the length of the period decreases.

Our paper is also related to work in mechanism design with limited commitment. If the principal discovers the agent is informed, he has the incentive to extract more information. This is similar to the "ratchet effect" facing a regulated firm which reveals it is efficient and is then punished by lower regulated prices or higher output in the future (we offer a discussion of the connections in Section 6). A principal's inability to commit can also dramatically affect incentives in a moral hazard setting. Padro i Miquel and Yared (2010) study a dynamic principal-agent model where jointly costly intervention is the only instrument the principal can utilize to give an agent incentives to exert effort. The principal must also be willing to carry out the punishment as there is limited commitment. Mialon, Mialon, and Stinchcombe (2012) study how the availability of torture as a mechanism creates commitment problems in other areas, specifically alternative counter-terrorism methods. They do not model the interrogation process or study the effectiveness of torture as a mechanism.

Lastly, the "deadline effect" in finite-horizon bargaining models with incomplete information (see Hart (1989), Ausubel and Deneckere (1992), and more recently Fuchs and Skrzypacz (2011)) bears some resemblance to our result that interrogation is delayed until near the terminal date. An important novelty in our model is that even in the absence of discounting this delay is costly to the principal because it limits the amount of time he has to extract valuable information from the suspect.

2 Model and Full Commitment

There is a torturer (principal) and a suspect (agent). There is a terrorist attack planned for a future date and the principal will try to extract as much information as possible prior to that date in order to avert the threat. Time is continuous and torture imposes a flow cost of Δ on the suspect. We assume that torture entails a flow cost to the principal of $c > 0$ so that torture will be used only if it is expected to yield valuable information.

The suspect might be *uninformed,* for example, a low value target with no useful intelligence about the terrorist attack, or an innocent bystander captured by mistake. On the other hand the suspect might be an *informed,* high value target with a quantity x of perfectly divisible, verifiable (i.e. "hard") information. The principal doesn't know which type of suspect he is holding and $\mu_0 \in (0, 1)$ is the prior probability that the suspect is informed.

We will consider two scenarios. In the *ticking time-bomb* scenario the attack is coming at a fixed known date. There is thus a finite time horizon T and no discounting.[11] If the suspect reveals the quantity $z \leq x$ and is tortured for time $\tau \leq T$, his payoff is

$$-z - \Delta\tau$$

while the principal's payoff in this case is

$$z - c\tau.$$

When the suspect is uninformed, z is necessarily equal to zero because the uninformed has no information to reveal.

In the *perpetual threat scenario*, the attack may come with a commonly known probability at any time. Thus the time horizon is infinite. In addition to discounting the future costs of torture, both parties discount the future payoffs from information revelation because the later the information is revealed the less likely it is to be useful in averting the threat.

With full commitment, torture gives rise to a mechanism design problem with verifiable information which is entirely standard except that there

[11] Incorporating discounting into the ticking time-bomb model would only complicate the notation without changing any of the qualitative results.

is no individual rationality constraint. Because the information is verifiable, the only incentive constraint is to dissuade the informed suspect from hiding his information. We simplify the exposition by using undiscounted payoffs in this full-commitment analysis.

The principal demands information $y \leq x$ from the suspect and commits to torture the agent for a duration $\tau(y)$ if the agent does not confess. If the informed suspect confesses, the principal's payoff is

$$y\mu_0 - (1-\mu_0)\, c\tau(y) \tag{1}$$

because he earns payoff y from the information revealed when the suspect is informed but must carry out his commitment to torture when the suspect is uninformed and has no information to reveal. The informed suspect optimally confesses whenever the cost of doing so is no larger than the cost of torture, i.e.

$$y \leq \Delta \cdot \tau(y). \tag{2}$$

Hence, the principal maximizes (1) subject to (2). The incentive constraint must bind at the optimum as the principal's payoff is increasing in the demand y. The optimal mechanism to induce revelation of y is to set $\tau(y) = y/\Delta$ and obtain payoff

$$y\left(\mu_0 - \frac{(1-\mu_0)\, c}{\Delta}\right).$$

If the term in parentheses is positive, the optimal mechanism is to demand that the suspect reveal the maximum amount of information. Otherwise, the principal demands zero.

The maximum amount of information that can be extracted depends on the length of the time horizon. The duration of torture required to induce the agent to reveal x is x/Δ. In the ticking time-bomb scenario, if this is longer than the time horizon T, then the time constraint binds and the principal can extract at most $y = \Delta T$. We summarize these full-commitment results below.

Theorem 1. *At the full commitment solution, if $\mu_0\Delta - (1-\mu_0)\, c \geq 0$, the principal demands information $\min\{x, T\Delta\}$ and inflicts torture for $\min\{x/\Delta, T\}$ periods if any less than this is given. If $\mu_0\Delta - (1-\mu_0)\, c < 0$, the principal does not demand any information and does not torture at all.*

3 Limited Commitment

In practice, torture takes place over time and the torturer has repeated opportunities to re-assess its execution. That is, the principal is not bound to any pre-committed plan of torture. If the suspect has not revealed anything, the principal may switch to a more promising prospect or may be under public pressure to stop torture. Hence, it is impossible to commit to continue to torture a suspect who stays silent. If the suspect is giving up actionable intelligence, the principal has every incentive to continue to extract more information. Stopping may even invite a backlash from the public. Hence, it is impossible to commit *not* to torture an agent who has given up information. In fact, with limited commitment, the principal's optimal policy at each point of re-assessment will be driven by the probability the agent is informed. The agent's optimal strategy of information revelation is in turn driven by the principal's strategy and we must study them in equilibrium.

To do this, we model limited commitment by dividing real time into periods of discrete time. We assume that the principal can only commit to torture for a single period. The form of commitment in a given period t is also limited. The principal can demand a (positive) quantity of information y_t and commit to suspend torture in the given period if the agent complies. In the event the agent does comply the principal's period-t payoff is $u_t = y_t$ and the agent's per-period payoff is $v_t = -y_t$. In the event that the agent refuses a positive demand from the principal, the principal and agent's period-t payoffs are $-c$ and $-\Delta$ respectively. Henceforth when we say the principal "tortures" the agent in period t we mean that he makes a non-zero demand $y_t > 0$. If the principal does not torture in period t, i.e. $y_t = 0$, then both parties earn a zero payoff in that period.

A pure strategy of the principal specifies for each past history of demands and revelations the choice of whether to threaten torture in the current period, and if so, what quantity $y \geq 0$ of information to demand. Note that a demand of $y = 0$ (which is the only demand that can be met by both the informed, costlessly, and uninformed suspect) is equivalent to pausing torture during the current period. For the informed agent, a pure strategy specifies for each past history and the present demand by the principal, the quantity of verifiable information to yield. The uninformed agent has no option but to reveal nothing every period.

We study the perfect Bayesian equilibria of this game.

3.1 The Perpetual Threat Scenario

The principal and agent discount future payoffs using the common discount factor δ, hence the principal's overall payoff is

$$(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}u_t$$

and the agent's is

$$(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}v_t.$$

The discounting reflects both the players' rate of time preference and the fact that the game may end with some exogenous probability in each period.[12]

In the perpetual threat scenario there is an infinite horizon and as is typical of such games there are multiple equilibria with varying payoffs for the principal. Indeed, no matter how likely it is that the suspect is informed, and no matter how much information he has conditional on being informed, there is always an equilibrium in which torture has zero value.

The structure of that worst-case equilibrium illustrates one of the central commitment problems inherent to torture and so it is worth describing it in detail. Consider what happens when the agent first *confesses*, i.e. the first period in which he concedes to a demand $y > 0$. The continuation game after a confession is one of complete information. In particular it is now common knowledge that the suspect is informed and moreover that he possesses additional information (i.e. $x - y$) that can be extracted. It is thus an equilibrium[13] of the continuation game for the principal to continue to torture and demand the remainder of the agent's information. The agent, anticipating the principal's resolve will optimally concede, i.e. accede to the principal's further demands after confession, until all of his information is revealed.

[12] If the game ends in some period, the agent's payoff is x and the principal's is $-x$ in that period and zero thereafter.

[13] It is of course not the only equilibrium. Indeed the amount of infomation that can be extracted in equilibrium of the complete information game ranges from zero to everything that remains.

In the worst-case equilibrium for the principal, the suspect refuses to confess precisely to avoid this "ratchet" effect. And because the suspect will always refuse, the principal's best-response is to never torture.

Proposition 1. *As long as $\Delta < \delta x$ (i.e. the period length is short enough) there exists an equilibrium in which the agent refuses any demands and the value of torture is zero.*

In this section we identify an upper bound to the value of torture across all equilibria. We begin by defining some quantities. Suppose that the agent has not yet confessed and let μ be the current probability that the suspect is informed and q the probability of confession in the current period. Then the posterior probability that the suspect is informed conditional on *not* revealing information is given by

$$B(q;\mu) = \frac{\mu(1-q)}{1-\mu q}. \tag{3}$$

The greater is the probability of confession q, the smaller is the probability the agent is informed if he does not confess.

A key observation is that despite the infinite horizon, any equilibrium necessarily has a last period of torture. To see why, note that as long as the suspect resists torture the probability that he is informed declines monotonically and therefore converges to some limit. Once the posterior is sufficiently close to that limit, the total probability of a confession in the remainder of the game (which of course is no larger than the probability that the suspect is informed) is so small that the expected value of any information extracted will be too low to justify the cost of torture. Therefore the principal stops.[14] The formal proof is in Appendix A.

Lemma 1. *There exists an integer n such that in any equilibrium, with probability 1, there are at most n periods in which the agent has not yet confessed and the principal demands $y > 0$.*

This lemma allows us to analyze the game as if it had a finite horizon, effectively characterizing equilibrium strategies by backward induction. Consider the informed suspect's incentives in the nth period of torture. By

[14] This sketch implicitly assumes the principal is using a pure strategy. Accounting for mixed strategies, as we do in the formal proof, is straightforward.

resisting he can avoid conceding any information since there is no threat of any future torture. Thus, his continuation payoff is at least $-\Delta$, the payoff from withstanding torture one last period. This means that if he were in fact to confess in this last period, then he cannot be induced to concede more than a quantity Δ. How can this be ensured? After all, in the worst-case equilibrium, once the agent confesses he then "spills his guts" and concedes all of x. But we can utilize the same expectation to stop the agent from confessing any further. In particular, after the initial concession of Δ, the continuation play will follow the strategies described in Proposition 1. In particular the principal optimally stops torturing because he knows that the agent will not agree to any further concessions because the agent knows that any further concession would lead him to spill his guts. The result is that an upper bound for the principal's continuation payoff starting from the nth period of torture is

$$\tilde{V}^1(\mu_1) = \Delta\mu_1 - c(1-\mu_1)$$

where μ_1 is the probability (conditional on having reached the nth period of torture with no prior confession) that the agent is informed.[15] In particular, the principal gets (at most) Δ from the informed agent, but must carry out his threat of torturing the uninformed agent who has nothing to concede.

The analysis proceeds by working backwards through the equilibrium strategies, bounding the principal's payoffs by an expression $\tilde{V}^k(\mu_k)$ which depends only on the number k of periods of torture remaining, and the conditional probability μ_k that the suspect is informed.

Let μ_2 be the posterior entering the second-to-last period in which the principal tortures.[16] Let q be the probability with which the informed suspect confesses in that period. Thus, the total probability of a confession is $\mu_2 q$. The principal's continuation payoff entering that period is bounded by

$$\mu_2 q 2\Delta + (1-\mu_2 q)\left[\tilde{V}^1(B(q;\mu_2)) - c\right].$$

To see why note that the suspect can secure a payoff of at least -2Δ from resisting torture for the two remaining periods (note that the loss in the final period of torture would be discounted.) Thus the principal can extract

[15] For ease of exposition, throughout this section we omit the normalizing factor $(1-\delta)$ in the expressions for payoff bounds.

[16] I.e. there have already been $n-2$ periods of torture and resistance.

at most 2Δ from a suspect who concedes in the second-to-last period of torture. In the event that the agent does not concede, the principal incurs the cost c and later obtains the continuation value from resuming torture which we have already bounded by $\tilde{V}^1(\cdot)$. As this payoff comes later, it would be discounted but that only lowers the principal's payoff even further.

The expression above embodies an important tradeoff. Recall that $B(q;\mu_2)$ is declining in q. If the suspect is induced to confess faster (formally, a larger q), then in the event he does not in fact confess the principal thinks it is more likely that the suspect is uninformed. This in turn lowers the continuation payoff $V^1(B(q;\mu_2))$ from any subsequent round of torture. Because the principal's strategy prescribes an additional period of torture later (and indeed this is necessary to induce the suspect to concede 2Δ) the continuation value to the principal from doing so must be non-negative, i.e. $\tilde{V}^1(B(q;\mu_2)) \geq 0$. This places an upper bound on the confession probability q.

This is the key difference between commitment and full commitment. With full commitment, the principal can credibly threaten to torture the agent in the future even when he knows he is uninformed. With limited commitment, the principal must believe that the agent is informed with a high enough probability to credibly threaten torture. This reduces the principal's payoff both because the informed suspect is less likely to confess under limited commitment and also because the principal ends up having to torture not only uninformed suspects but informed suspects who do not confess. That is, the following constrained maximization represents an upper bound on the principal's payoff entering the second-to-last period in which he tortures.

$$\max_q \mu_2 q 2\Delta + (1-\mu_2 q)\left[\tilde{V}^1(B(q;\mu_2)) - c\right] \tag{4}$$

$$\text{such that } \tilde{V}^1(B(q;\mu_2)) \geq 0 \tag{5}$$

The principal gains at most 2Δ if the agent confesses in the second-to-last period but gets at most $\Delta - c$ if he confesses in the last period. So, earlier confession increases total information conceded and saves on the cost of torture. Therefore, the maximand is increasing in the probability of confession and, since $\tilde{V}^1(B(q;\mu_2))$ is strictly decreasing in q, the constraint binds. Thus, if we define the maximal confession probability $\tilde{q}_2(\mu_2)$ by the equation

$$\tilde{V}^1(B(\tilde{q}_2(\mu_2), \mu_2)) = 0$$

then the maximum is achieved by $\tilde{q}_2(\mu_2)$ and thus the principal's payoff is bounded by $\tilde{V}^2(\mu_2)$ defined as follows

$$\tilde{V}^2(\mu_2) = \mu_2\tilde{q}_2(\mu_2)2\Delta - (1 - \mu_2\tilde{q}_2(\mu_2))c.$$

Continuing in this fashion we can inductively define sequence of functions $\tilde{V}^k(\mu)$ and $\tilde{q}_k(\mu)$ and probabilities $\tilde{\mu}_k$ as follows.

$$\tilde{V}^k(\mu) = \mu\tilde{q}_k(\mu)k\Delta - c(1 - \mu\tilde{q}_k(\mu)), \tag{6}$$

$$\tilde{V}^k(\tilde{\mu}_k) = 0, \tag{7}$$

$$B(\tilde{q}_k(\mu); \mu) = \tilde{\mu}_{k-1}. \tag{8}$$

With k periods of torture remaining the probability that the agent is informed must be sufficiently high in order for the principal's expected payoff to be high enough to make him willing to carry on torturing those k additional periods. The cutoff $\tilde{\mu}_k$, defined in Equation 7, represents that minimum probability. There is an upper bound to how quickly the suspect can be induced to confess without the conditional probability falling below $\tilde{\mu}_k$. Equation 8 defines the corresponding maximum rate of confession $\tilde{q}_k(\mu)$ as a function of the current probability μ that the suspect is informed. Finally, given these bounds, we can compute a bound on the principal's payoff from torturing for k additional periods, given in Equation 6.

In particular since the principal's strategy prescribes a total number of periods of torture no greater than n, the principal's equilibrium payoff is no greater than $\tilde{V}^n(\mu_0)$.

Lemma 2. *Let n be the maximum number of periods the principal is willing to torture. Then the value of torture is no greater than $\tilde{V}^n(\mu_0)$.*

Next we can strengthen the bound by combining Lemma 1 and Lemma 2. In particular we can characterize, as a function of the prior μ_0, the maximum number of torture periods n. Whereas we obtained the bound in Lemma 2 by arguing that there was a *maximum* rate of confession q, we can derive further implications from the observation that there is also a *minimum* rate of confession in equilibrium. To see why, note that the principal's payoff from torturing in a given period is positive only if he expects the agent to confess with at least some minimal probability. If the confession probability is too low, then with high probability the principal will be incurring the cost of torture for too little gain. Moreover, this minimal confession rate is lower the earlier the principal begins torture. Otherwise, if the agent does *not* confess, the principal puts too low a probability on the suspect being informed to subject him to torture for the time that remains. Given the rate of confession is low if torture begins early, the principal will only begin if his prior is high. This identifies a bound on the maximum number of periods the principal will torture. This argument underlies the main result of this section.

Theorem 2. *Fix the prior μ_0 and let $K(\mu_0)$ to be the largest k such that the sum*

$$\sum_{j=1}^{k}(1-\mu_0)\left[\frac{c}{j\Delta + c}\right]$$

is no larger than μ_0.

1. *Regardless of the value of x, the principal tortures for at most $K(\mu_0)$ periods.*

2. *Regardless of the value of x, the principal's payoff is less than*

$$\max_{k \leq K(\mu_0)} \tilde{V}^k(\mu_0).$$

3. *In particular, the value of torture is bounded by*

$$K(\mu_0)\Delta$$

The see the significance of the bound given in Theorem 2, note that $K(\mu_0)$ is independent of x. That is, no matter how much (or how valuable) is the information held by the informed suspect, there is a fixed upper bound on the number of periods in which the principal can credibly threaten him with torture. Since the informed suspect will suffer at most $K(\mu_0)\Delta$ under torture, this places a corresponding limit on the total amount of information that can be extracted, independent of how much information is held.

The bound in Theorem 2 applies to the model in which the principal is able to commit to torture and impose a flow cost Δ for a discrete time period. Because of the inherent commitment problems, the discrete nature of these torture episodes helps the principal and inflates the value of torture. To further emphasize the limitations of torture as a mechanism for extracting information, we will later consider shortening the time interval between opportunities to continue torturing. Shortening the period length reduces both the threat and the cost to the principal that a period of torture represents. It also gives the principal multiple opportunities to revisit his torture policy in any given length of physical time and therefore magnifies his commitment problem. In fact, we will show (see Theorem 5) that these extra opportunities to torture imply the constant $K(\mu_0)$ is independent of the period length. Thus, without the commitment effect that discrete time entails, the value of torture vanishes.

3.2 The Ticking Time-Bomb Scenario

The classical argument for torture imagines there is a terrorist attack set to take place at a known date unless information is obtained that helps to prevent the attack. Proponents of torture invoke this situation presumably because they believe it represents the best case for torture[17]. Hence, any strategic analysis of torture must deal with the ticking time-bomb scenario which we turn to in this section.

In the ticking time-bomb scenario there is a finite horizon, i.e. a fixed number T discrete time periods after which the game ends. We consider undiscounted payoffs for notational simplicity, adding discounting would not change any of the results. Thus if the suspect concedes a total amount of information z and withstands torture in k periods then his payoff is $-z - k\Delta$ and the payoff to the principal is $z - kc$. We measure time in reverse, so "period k" means that there are k periods remaining. But "the first period" and "the last period" mean what they usually do.

We begin with a series of observations that accentuate the comparison between the ticking time-bomb and perpetual threat scenarios. In both scenarios, once the suspect reveals some information, say in period k, the continuation game is one of complete information. In the perpetual threat scenario we showed that this complete information subgame has multiple equilibria. By contrast, in the ticking time-bomb scenario the continuation equilibrium is essentially unique. As shown in the following lemma, in all

[17] See for example Dershowitz (2002) and Leibovich (2009).

equilibria of the continuation game beginning in period $k-1$, the suspect "spills his guts," i.e. he reveals all of his remaining information, up to the maximum he can be induced to reveal. That maximum is given by the total remaining costs of torture that can be threatened: $(k-1)\Delta$, in other words Δ per period. He cannot reveal this information in one period because the principal would then continue to extract more information from him in the time that remains.[18] The proof is via backward induction and can be found in Section B.2.

Proposition 2. *In any equilibrium, at the beginning of the complete information continuation game with k periods remaining and a quantity $z > 0$ of information yet to be revealed, the suspect will be induced to concede the remainder of his information and have continuation payoff*

$$-\min\{z, k\Delta\}$$

Thus, an informed agent faces a large punishment as soon as he confesses. He is therefore only willing to confess if he expects to face an equivalent threat were he to reveal nothing. That is, the principal must also be expected to continue torturing a suspect who reveals nothing. Indeed, this logic implies that once the torture begins, it cannot stop. This is because in each period the principal tortures he must be expected to continuing torturing if the suspect resists. We formalize this in the following proposition.

Define $\bar{k}$ to be the largest integer strictly smaller than x/Δ. Thus, $\bar{k}+1$ measures the minimum number of periods the principal must be prepared to torture in order to induce revelation of the quantity x. We will refer to the phase of the game in which there are $\bar{k}$ or fewer periods remaining as the *ticking time-bomb* phase. In the ticking time-bomb phase, the limited time remaining is a binding constraint on the amount of information that can be extracted through torture. Next, we say that there is effective torture in period k if the principal makes a positive demand and the suspect concedes with positive probability.

[18] Horner and Skrzypacz (2015) study a signaling model where a "competent" agent with information that is without value to him but is of value to a principal attempts to separate from an uninformed "incompetent" agent. By releasing information slowly, the informed agent can separate himself from the uninformed more easily as the uninformed has more chances to fail to mimic the informed. In our paper, information is released slowly to prevent the ratchet effect that arises as the principal canot stop himself from demanding as much as information as possible in the time available.

Proposition 3. *Within the ticking time-bomb phase, once effective torture begins it must continue uninterrupted until the end.*

This has two important implications which limit the value of torture. First, almost all of the effective torture must happen near the deadline and within the ticking time-bomb phase. To see why, suppose that there is effective torture in period k earlier than the ticking time-bomb phase. Then by Proposition 2 the payoff to a suspect who confesses is $-x$. This must be at least as large as the payoff to a suspect who resists. Therefore a suspect who resists must be tortured for an additional $(x/\Delta) - 1$ periods.[19] Not one of those additional periods of torture can occur prior to the ticking time-bomb phase. Because if so then by the same argument there must be an additional $(x/\Delta) - 1$ periods of torture after that implying that the resistant suspect faces torture for at least $x/\Delta + 1$ periods and therefore has payoff no larger than $-\Delta\left[(x/\Delta) + 1\right] < -x$. This is impossible since the suspect would rather confess and have payoff $-x$. But then, since an informed suspect confesses with probability one, the principal will not torture a suspect who resists as he is known to be uninformed, a contradiction.[20]

Proposition 4. *There can be at most a single period of effective torture prior to the ticking time-bomb phase.*

Secondly, even within the ticking time-bomb phase Proposition 3 implies a bound on how early effective torture can begin. As in the perpetual threat scenario, there is a minimal rate of confession very period, otherwise the principal finds the costs of torture greater than the benefits. In the ticking time-bomb scenario, this rate of confession and in fact the unique equilibrium is characterized via a backward induction argument.

We have already shown in Proposition 2 what happens after confession. The remainder of the analysis focuses on the behavior along a path

[19] Ignoring integer issues for this heuristic argument.

[20] It is worth noting that Proposition 4 also holds under discounting. This is in turn because Proposition 2 also holds under discounting. The suspect concedes Δ units of information per period as the maximum cost the principal can impose each period is Δ and slowing down the release of information increases the suspect's discounted payoffs. But he does spill his guts. Hence, the principal must torture a suspect who resists for an equivalent amount of time. But then the argument in the text obtains and there is at most one period of effective torture outside the ticking time-bomb phase.

in which the informed suspect resists. Beginning with the final period, period 1, define

$$V^1(\mu) = \Delta\mu - c(1-\mu).$$

The function V^1 represents the principal's unique equilibrium continuation payoff when μ is the conditional probability that the (heretofore resistant) suspect is informed. To characterize behavior in earlier periods we next define μ_1^* by

$$V^1(\mu_1^*) = 0.$$

A posterior μ_1^* makes the principal indifferent between torturing or not in period 1. This condition will pin down the probability of confession in period 2. Define $q_2(\mu)$ as the solution to the following equation.

$$B(q_2(\mu); \mu) = \mu_1^*$$

i.e.

$$q_2(\mu) = \frac{\mu - \mu_1^*}{\mu(1-\mu_1^*)}.$$

Suppose the suspect has kept silent up to period 2 and μ is the probability he is informed. Then by confessing in period 2 with probability $q_2(\mu)$, he insures that, in the $1 - q_2(\mu)$-probability event that he does not confess, the principal is indifferent between torturing or not in the final period. Thus $q_2(\mu)$ is the maximum equilibrium confession rate in period 2: any larger confession probability would leave the principal unwilling to continue torturing a suspect who resists violating Proposition 3. We show in Appendix B that in equilibrium the suspect must be conceding at this maximal rate otherwise the principal can slightly reduce his demand and induce the agent to concede faster.

To extend the analysis to earlier periods, we inductively define functions $V^k(\mu)$ and $q_k(\mu)$ and probabilities μ_k^*. In the essentially unique equilibrium $V^k(\mu)$ and $q_k(\mu)$ are the value of torture and the probability of confession when k periods of effective torture remain and the suspect is informed with probability μ. We show in Appendix B that these quantities are well-defined.

$$V^k(\mu) = \mu q_k(\mu)\min\{x, k\Delta\} + (1-\mu q_k(\mu))\left[V^{k-1}(\mu_{k-1}^*) - c\right]. \tag{9}$$

$$V^k(\mu_k^*) = V^{k-1}(\mu_k^*) \tag{10}$$

$$B(q_k(\mu); \mu) = \mu_{k-1}^*. \tag{11}$$

Theorem 3. *In the ticking time-bomb scenario the equilibrium is unique up to payoff-irrelevant variations. The unique equilibrium payoff for the principal is*

$$\max_{k \leq \bar{k}+1} V^k(\mu_0).$$

The essentially unique equilibrium of the game has the following path of play. The principal chooses the k^* which achieves the maximum continuation value above and begins torturing in that period. In each period of torture he demands Δ. If ever the suspect confesses he then concedes the maximum amount of information according to Proposition 2. As we will show later, typically $k^* < \bar{k}+1$, i.e. the principal waits for the ticking time-bomb phase before commencing torture.[21] In such cases, in accordance with Proposition 3 the principal tortures a resistant agent with probability 1 in all remaining periods k. In the first period of torture the suspect confesses with probability $q_{k^*}(\mu_0)$. This ensures that, conditional on no confession the updated probability he is informed will be $\mu^*_{k^*-1}$, i.e. in the next period the principal will be (just) willing to continue torturing. In all subsequent periods k the updated posterior will be μ^*_k and the heretofore resistant suspect will confess with probability $q_k(\mu^*_k)$. In the final period the agent confesses with probability 1 if he had not confessed previously.[22]

3.3 Comparing The Two Scenarios

A comparison of the perpetual threat and ticking time-bomb scenarios sheds further light on value of torture with commitment problems. The urgency of the ticking time-bomb might be thought to strengthen the resolve of the principal and hence makes this scenario the leading case in favor of torture as an information extraction mechanism. But we show that the perpetual threat scenario actually makes the best case for torture.

The ticking time-bomb scenario does at least put a positive lower bound on the value of torture to the principal. At dates close enough to the ticking time-bomb the suspect knows that if he confesses today there is limited time to extract the remaining information. This acts as substitute for

[21] Indeed, as implied by Proposition 4 even in exceptional cases all but one period of torture occurs within the ticking time-bomb phase.

[22] In Appendix B, the complete description of equilibrium strategies is given, including off-path beliefs and behavior and we prove Theorem 3.

a credible promise on the part of the principal that in exchange for a small amount of information today the torture will stop soon. At worst the principal can always wait until very near the ticking time-bomb to begin the torture, effectively sacrificing the ability to extract a lot of information in exchange for a guarantee that he extracts at least some information. This is one, perhaps less obvious, argument in support of the usual position that the ticking time-bomb makes the strongest case for torture.

Absent a fixed and known last opportunity to torture the principal has no credible commitment to stop torturing an informed suspect. An informed suspect therefore rationally anticipates that even the smallest initial confession will eventually result in further torture and extraction of additional information. This "ratchet" effect can dissuade the suspect from conceding even in the first instance and this logic underlies the zero value of torture in worst-case equilibrium in the perpetual threat scenario.

However, this effect of the time horizon has a flipside. As we discussed, the no-torture equilibrium described in Proposition 1 can also serve as a *continuation equilibrium* after any history of torture signifying the (commonly anticipated) end of further interrogation. Most importantly the time at which torture ends is determined only by expectations and therefore not tied to any arbitrary deadline. By contrast, in the ticking time-bomb scenario the only way to capitalize on the commitment power of the torture deadline is to wait long enough to begin torturing.

Thus, with discounting, the maximum equilibrium value or torture if higher in the perpetual threat scenario than the unique equilibrium value from the ticking-bomb scenario. A simple way to demonstrate this is to take the ticking time-bomb equilibrium and simply move the initial date of torture to the very first period of the game leaving the total duration of torture and all other aspects of the strategies otherwise unchanged. The end-date of torture is now enforced not by the deadline but by inserting Proposition 1 as the continuation equilibrium at that date. It is easy to see that this constitutes an equilibrium of the perpetual threat scenario which extracts the same amount of information, only earlier.

In fact, the best equilibrium of the perpetual threat scenario is even better than this and the reason stems from the relative impact of the second commitment problem in the two contexts. The principal benefits from a commitment to *continue* torturing a suspect who has yet to confess. Indeed our analysis of the full commitment solution utilizes this form of commitment to induce an informed suspect to confess immediately rather than face certain lengthy torture. More generally, the longer the principal can be expected to torture a resistant suspect the stronger is his incentive to confess early.

The friction that limits the duration of torture is sequential rationality for the principal: the continuation value of torture must be high enough to justify it. But the minimum continuation value necessary is lower in the perpetual threat scenario than in the ticking time-bomb. We can see this through a comparison of the conditions that define equilibrium in the two cases. Consider Figure 1. It shows the value functions $V^k(\mu)$ for the ticking time-bomb scenario that encode the value of torturing for k periods as a function of μ. We know from the unique equilibrium in that scenario that if the prior μ_0 exceeds μ_2^* then the probability of confession will be just high enough so that when period 2 arrives the posterior will be μ_2^*.

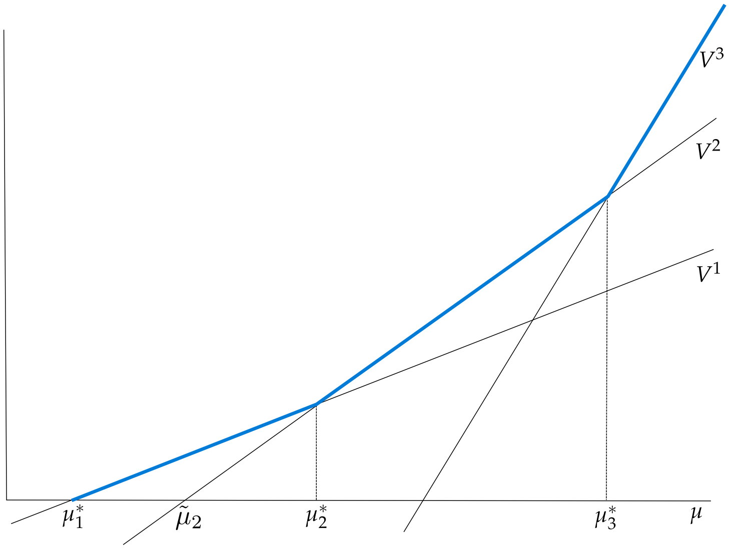

Figure 1: An illustration of the functions V^k and the thresholds μ_k^*. Here $\bar{k} + 1 = 3$. The upper envelope shows the value of torture as a function of the prior μ_0.

The posterior should be μ_2^* because this posterior equates the value of continuing to torture for the remaining 2 periods with the value of pausing for one period and torturing only in the last period. In other words $V^2(\mu_2^*) = V^1(\mu_2^*)$, as in Equation 10. If the suspect had conceded with any higher probability then the posterior would have fallen below μ_2^* and the principal would not carry out his expected continuation strategy of torturing uninterrupted until the end. But then the suspect will not confess in period 3 as he will not be tortured in period 2 if he resists. This is why Equation 10 must be satisfied in equilibrium in the ticking time-bomb scenario.

In the perpetual-threat scenario, by contrast, the necessary condition is given by Equation 7, in particular the posterior should be $\tilde{\mu}_2$ which is defined by[23] $V^2(\tilde{\mu}_2) = 0$. This is because if the principal skips a period of torture, he can be punished with the no-torture, zero value equilibrium. This makes it easier to satisfy the principal's sequential rationality constraint. One can see from the figure that $\tilde{\mu}_2 < \mu_2^*$ and thus the rate of confession is higher in the perpetual threat scenario. Indeed this comparison holds for all earlier periods as shown in the following proposition.

Theorem 4. *For all k, and for all μ,*

1. $\tilde{q}_k(\mu) \geq q_k(\mu)$
2. $\tilde{V}^k(\mu) \geq V^k(\mu)$

with a strict inequality for $k \geq 3$. Moreover, as a consequence the number of periods of effective torture in the ticking time-bomb scenario is bounded by $K(\mu_0)$ just as in the perpetual threat scenario.

An implication of the theorem is that the best equilibrium in the perpetual threat scenario improves over the ticking time-bomb not just because torture can be started without delay but also because confessions can be induced with a higher probability even over the same number of periods, increasing the amount of information extracted and reducing the costs of carrying out torture. Also, if intensifying commitment problems diminish the value of torture in the perpetual threat scenario, they must also diminish the value in the ticking time-bomb scenario. We turn to this issue in the next section.

[23] To be precise, Equation 7 reads $\tilde{V}^2(\tilde{\mu}_2) = 0$, but it is easy to see by a comparison of Equation 9 and Equation 6 that $\tilde{V}^2 \equiv V^2$, and indeed we show this in the proof of Theorem 4.

4 Shortening The Period Length

We model the principal's limited commitment by allowing repeated opportunities to revisit whether to continue torturing. In practice the principal may be able to revisit his strategy almost continuously, further reducing his power to commit. To what extent is the value of torture dependent on the ability to commit to carry out torture over a discrete period of time? To answer this question we now consider a model in which the period length is parameterized by $l > 0$. The model analyzed until now corresponds to the benchmark in which $l = 1$. We study the value of torture to the principal as the period length shrinks.

The torture technology is parameterized by its continuous-time flow cost to the suspect (Δ) and to the principal (c.) Translated into discrete time, when the period length is l, the total cost of a single period of torture is

$$\Delta_l = \int_0^l r\Delta e^{-rs} ds = (1 - \delta^l)\Delta \tag{12}$$

for the agent and

$$c_l = \int_0^l rce^{-rs} ds = (1 - \delta^l)c \tag{13}$$

for the principal, where $r > 0$ is the continuous-time discount rate and $\delta = e^{-r}$. Similarly, in the ticking time-bomb model (without discounting) these costs are $\Delta_l = l \cdot \Delta$ and $c_l = l \cdot c$ respectively.

With these as discrete-time payoffs we can apply the results in Theorem 2 and Theorem 4 to bound the value of torture in both the perpetual threat and ticking time-bomb scenarios as a function of the period length l. The system of equations that defines the bounds $\tilde{V}^k(\mu)$ is reproduced below, now parameterized by l.

$$\tilde{V}^{k,l}(\mu) = \mu\tilde{q}_{k,l}(\mu)k\Delta_l - c_l(1 - \mu\tilde{q}_{k,l}(\mu)).$$
$$\tilde{V}^{k,l}(\tilde{\mu}_{k,l}) = 0$$
$$B(\tilde{q}_{k,l}(\mu); \mu) = \tilde{\mu}_{k-1,l}.$$

From Equation 12 and Equation 13 (and the corresponding expressions for $\tilde{V}^{k,l}(\mu)$ for the ticking time-bomb model) we see that

$$\tilde{V}^{k,l}(\mu) = Z_l \tilde{V}^{k,1}(\mu)$$

where Z_l is independent of μ and

$$\lim_{l \to 0} Z_l = 0.$$

Thus, for every k, the thresholds $\tilde{\mu}_{k,l}$ and furthermore the concession rates $\tilde{q}_{k,l}(\mu)$ are independent of l.

It follows from Theorem 4 and Theorem 2 that $Z_l \tilde{V}^k(\mu)$ is an upper bound on the principal's continuation payoff when there are k periods of torture remaining and the period length is l. Moreover regardless of the period length, $K(\mu_0)$ is an upper bound on the number of periods of torture and $Z_l K(\mu_0)$ is therefore an upper bound on the real-time duration of effective torture. In particular, the value of torture is bounded by

$$Z_l \Delta K(\mu_0).$$

Noting that the constant $K(\mu_0)$ depends only on the the prior μ_0 and the flow costs of torture c and Δ we have established the following.

Theorem 5. *In the limit as the time interval between decisions to continue torture approaches zero, the value of torture is zero.*

The ultimate source of the value of torture is the temporal commitment power given by discrete torture episodes. When these discrete periods are short, there are more points where the principal can stop torture in any given physical length of time. For the principal to continue to torture a suspect who has not confessed in the last half say of any length of time, he must put high probability on the suspect being informed as there are more points later on where he can stop and, if he does, the whole equilibrium would unravel. But this then implies the suspect must be confessing slowly in the first half of this length of time, so slowly that is not worth torturing him for information. By contrast, if there were only two periods in this length of physical time, the principal would be willing to torture in the last period even if the probability the suspect is informed is lower as the principal can commit for longer. But this means the rate of confession can be higher in the first half of time so the principal is willing to torture. This logic actually implies the principal will torture for a vanishing length of time as the period length goes to zero and hence can induce revelation of only a vanishing amount of information.

5 Extensions

We explore some natural extensions of the basic model. We allow information to be indivisible and the principal's payoff to be convex in information extracted. We allow the principal to choose between different interrogation technologies in each period. Finally, we explore the use of "carrots" rather than "sticks" by assuming torture can be made illegal and only monetary payments can be used to persuade the agent to reveal information.

5.1 Divisibility of Information

We have assumed that the information held by the informed detainee is a perfectly divisible quantity x and that the value to the principal of acquiring any portion $y \leq x$ is linear and equal to y itself. We can generalize this model by supposing that the value of an quantity y of information is given by some increasing function $v(y)$ where $v(0) = 0$ and $v(x) = x$ (the latter being normalizations which maintain as much consistency as possible with the preceding analysis).

For example, it might be natural under some interpretations to assume that $v(y)$ is convex. This would model a situation in which multiple pieces of information have complementary value. An extreme example would be where x represents the combination required to defuse the ticking time-bomb. Knowing anything less than the full combination would be of zero value to the Principal and therefore $v(\cdot)$ would be a step function where

$$v(y) = \begin{cases} 0 & \text{if } y < x \\ x & \text{otherwise} \end{cases}$$

More generally, we may take $w(y)$ to be some increasing function representing the probability that the attack can be averted when the principal has extracted the quantity y of information, and set

$$v(y) = x \cdot w(y)$$

so that x is the value of averting the attack. Then the principal's payoff from extracting y and torturing for a length of time t is

$$v(y) - ct$$

while the agent's is

$$-v(y) - \Delta t$$

(for simplicity we ignore discounting).

Regardless of the interpretation, or the details, as long as $v(\cdot)$ is a continuous function, all of the preceding analysis goes through unchanged when we simply re-normalize the units in which information is measured. In particular the principal demands information in units of $v(\cdot)$. Initially the principal demands $y = v^{-1}(\Delta)$, then proceeds by extracting pieces whose incremental value is Δ, i.e. next $v^{-1}(2\Delta) - v^{-1}(\Delta)$, then $v^{-1}(3\Delta) - v^{-1}(2\Delta)$, etc.

A continuous v represents information whose value is not linear in the quantity but which is nevertheless infinitely divisible. Divisibility of information only helps the principal because it enables him to fine-tune his demands in order to maintain incentives for the agent to confess. To see this, consider now the case of perfectly indivisible information where v takes the step-function form given above. In this case, once the time-period is short enough so that $\Delta < x$, the equilibrium has zero information revealed and therefore no torture at all.

To see why, consider the last period of the game and suppose the agent has thus far conceded $y < x$ to the principal. The agent can refuse any demand and secure a continuation payoff of at least $-\Delta$ by withstanding the last period of torture. In order for the principal to obtain a non-negative payoff he must demand the entire remaining quantity of information since any less has zero value. But since such a concession gives the agent $-x < -\Delta$ he would refuse.

In equilibrium, there will be no torture in the last period of the game no matter how much information has been conceded previously. By induction then there will be no torture in the penultimate period or in any period at all. To summarize:

Theorem 6. *Suppose the value of information is continuous. As the time interval between decisions to continue torture approaches zero, the value of torture is zero. Suppose the value of information takes the step-function form. Then the value of torture is zero as long as the period length is small enough to guarantee that* $x > \Delta$.

5.2 Enhanced Interrogation Techniques And The Ratchet Effect

Up to now, we have taken the torture technology as given. Instead suppose the principal has a choice of torture instruments, including a harsh enhanced interrogation technique. Perhaps the technology was considered illegal before and legal experts now decide that its use does not violate the letter of the law. Or in a time of war, norms of acceptable torture practices are relaxed. Enhanced interrogation techniques increase both the information that can be extracted every period and the cost to the principal. For example, sleep deprivation is less costly both to the suspect and the principal than waterboarding.

This creates another potential commitment problem for the principal - he might deviate and switch interrogation techniques in midstream. In the perpetual threat scenario, this issue does not arise as the no-torture equilibrium can be used to punish a deviation by the principal. But in the ticking time-bomb scenario, this second commitment problem does impact the principal's welfare. We can see this is in a simple two period example.

Let (Δ', c') denote the cost to the suspect and principal from the harsher technology. A tradeoff arises when the enhanced threat $\Delta' > \Delta$ comes at the expense of a more-than-proportional increase in the cost to the principal: $c'/\Delta' > c/\Delta$. In that case, the relative effectiveness of the two methods will depend on the the principal's prior. The more likely the suspect is to be uninformed, the better it is for the principal to use the milder technology as the chances of actually using it on the equilibrium path are higher. This can be seen in a simple example illustrated in Figure 2.

In the figure we have plotted the upper envelope of the V^k functions for the milder technology in bold (blue). The function V^1 for the harsher technology with a dashed (red) line. The relative positions of the two values of μ_1^* follows from the definition

$$\mu_1^* = \frac{c}{\Delta + c}.$$

As can be seen from the figure, for low priors μ_0, the principal prefers to use the milder technology for multiple periods whereas for greater priors the principal prefers to take advantage of the harsher technology and torture for fewer periods.

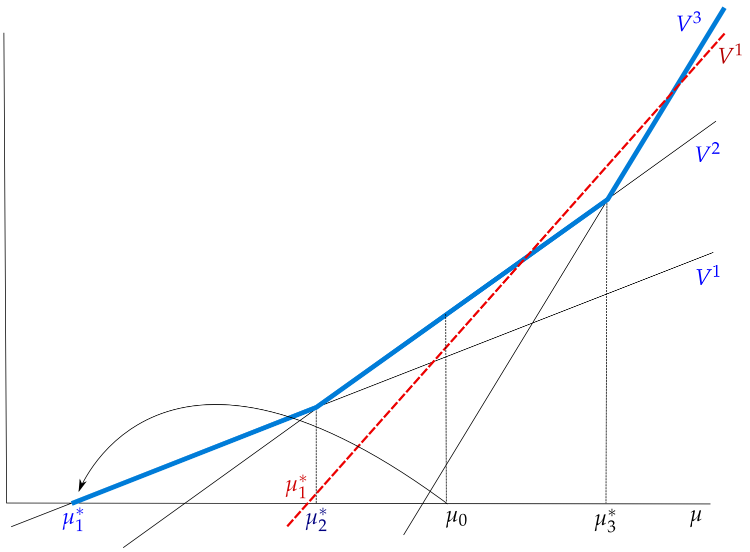

Figure 2: Enhanced interrogation methods undermine the principal's commitment power.

However, importantly it does not follow that the principal benefits from an array of technologies from which to choose depending on the context. The equilibrium where the principal uses the milder technology is predicated on his commitment to use that same technology for the duration. Making the harsher technology available comes at a cost even when the principal prefers not to use it because it can undermine this commitment.

To illustrate, refer again to figure Figure 2. Suppose that the prior probability of an informed suspect is μ_0. In this case the value of torture is maximized by using the milder technology for 2 periods. Consider how the corresponding equilibrium will unfold. In the first period of torture, the principal demands the quantity of information $y = \Delta$. The informed suspect expects that by yielding Δ, he will reveal himself to be informed and be forced to give an additional Δ in the final period. He accepts this because he knows that his payoff would be the same if he were to refuse: he will incur a cost of torture Δ in the current period and then accept the principal's demand of Δ in the last period.

But if the enhanced interrogation technique is available, this equilibrium unravels. Once the suspect reveals himself to be informed in period

2, the principal will then switch to the harsher technology for the last period in order to extract an additional Δ' from the suspect. This means that the suspect's payoff from yielding in period 2 is $-(\Delta + \Delta'.)$ On the other hand, if the suspect resists in period 2, his payoff remains -2Δ. This can be seen from Figure 2. In equilibrium after resistance in period 2 the posterior moves to the left to μ_1^* and the principal will optimally continue with the milder technology.

This commitment problem arises due to the ratchet effect. The principal benefits from a commitment to a milder technology. This allows him to convince the informed suspect that torture will be limited. However, once the suspect has revealed himself to be informed, the principal's incentive to ratchet-up the torture increases. When the enhanced interrogation method is available the principal cannot commit not to use it and his preferred equilibrium unravels. Indeed, without a commitment not to use the harsher technology, the equilibrium will be worse for the principal. The suspect will refuse any demand in the first period and the principal will be forced to wait until the last period and use the harsher technology.

5.3 Monetary Payments

Torture is a "stick" that can be used as a threat to punish a suspect who does not concede information. Monetary payments can be instead be used as a "carrot" to reward a suspect who does concede information. Of course, to citizens, paying for information might seem as abhorrent as torture. Monetary rewards also create perverse incentive effects and encourages crime. Setting these objections aside, suppose payments are allowed but are also subject to lack of commitment - the principal can renege on future payments perhaps because of the political difficulties associated with payment.

If the principal faces a choice between payments and torture, once the agent starts talking and the principal knows he is informed, the principal's trade-off changes and he favors torture over payments. This is because he will never actually use torture on the equilibrium path so it is costless while transfers are costly. Now the agent faces a ratchet effect if he talks because instead of getting a carrot of a monetary transfer to compensate him for giving up information, he faces a stick. This in turn implies he must be tortured if he does not confess, otherwise he will never talk. So, the possibility of payments does not eliminate the costs or the use of torture.

If all forms of torture or costly interrogation can be made illegal, monetary payments are the only instrument for information extraction and the ratchet effect does not arise. But the principal still faces a commitment problem because he can renege on payments so the value of interrogation is still not clear. To investigate this formally, suppose for payment p and information y suppose the principal's payoff is

$$v(y) - p$$

and the agent's is

$$-y + p.$$

To ensure there are gains from trade, we assume surplus $v(y) - y$ is maximized at $y^* > 0$ which for simplicity we take to be unique.[24]

For brevity, we focus on the ticking time-bomb scenario. Suppose the period length is l and the principal credibly hand over at most $l\Delta > 0$ in money each period. In particular, in each period the principal can make a take-it-or-leave-it offer (p, q) where he transfers $p \in [0, l\Delta]$ if and only if he receives at least q units of information. We use this construction to capture the idea that the principal can renege on monetary transfers and can credibly hand over at most $l\Delta$ each period. The informed agent releases no information and the informed agent releases $y \geq 0$ up to the information he has left.

Consider the following strategies. With t periods to go, let $x'(t)$ be the information the suspect has conceded. If $y^* > x'(t)$, the principal's demand q is the minimum of $l\Delta$ and $y^* - x'(t)$. If $y^* \leq x'(t)$, the principal's demand is $q = 0$. The principal offers the transfer $p = q$. Facing an offer (p', q'), the suspect accepts iff $p' - q' \geq 0$.

Given the principal's strategy, the informed suspect's continuation payoff is constant whether he accepts or rejects an offer in any given period. Hence, the informed suspect's strategy to accept the principal's offer if and only if $p' - q' \geq 0$ is sequentially rational. If the principal deviates in any period to an offer the informed suspect accepts, he is worse off as he must be giving the suspect positive surplus. If the principal deviates

[24] One natural payoff function has $v(y) = \lambda y$ where $\lambda > 1$. By rescaling the principal's payoff to $y - \frac{c}{\lambda} t$ we see that our formal results on the use of torture also apply to this model.

to an offer the suspect rejects, he is either no better off or is worse off if there is not enough time to extract the suspect's information before the ticking time-bomb explodes. Therefore, the principal's strategy is also sequentially rational and the strategies are a perfect Bayesian equilibrium. We have the following:

Theorem 7. *Suppose torture can be outlawed and monetary payments are feasible. Then, there is an equilibrium where the principal's payoff is first-best whatever the period length.*

The costs imposed on the principal by torture and monetary payments differ and hence create different implications for commitment. It is costless to offer the uninformed suspect a transfer if he confesses: he never confesses and the principal does not pay. Also, there is no ratchet effect after the suspect confesses in return for a transfer as further information extraction is costly for the principal because it requires further transfers. So commitment problems do not undermine monetary payments as a tool of information extraction when torture can be outlawed.

6 Difficulties with Commitment

If the principal can commit to torture a suspect even when he is certain the suspect is uninformed, he can implement the second-best solution identified in Theorem 1. The classical solution is a contract which specifies a verifiable action by the principal as a function of a verifiable report by the agent. The agent escapes torture if and only if he releases the information the principal demands. There is a third party, "the court", that enforces the contract and imposes a punitive fine on the principal should he deviate from the prescription of the contract. Alternatively, the full commitment solution can be implemented in a repeated game. Suppose the principal faces torture environments repeatedly, facing a different agent in each environment. If the principal deviates from the commitment solution with one agent, he loses his reputation and is punished by a switch to a punishment phase in future interactions. A sufficiently patient principal does not deviate. Both implementations face significant hurdles in the torture environment.

The contracting implementation is difficult even in economic environments. Contracting parties can renegotiate to a better allocation or the principal can renege and ratchet up incentives.[25] The same incentives arise in the torture environment and are compounded by another feature of the torture environment: Torture is carried out in secret so it is impossible to determine if the principal deviated from the terms of the contract or not. The terms of trade are verifiable in the buyer-seller setting but unobservable principal moral hazard undermines the optimal contract in the torture environment. The same issue compromises the implementation of the optimal contract via a repeated game. Players in future interactions with the principal cannot know whether the principal deviated from the optimal contract in the past with another player. Making torture verifiable does not help. Suppose the principal faces re-election motives and hence inculcates the preferences of the median voter. If torture is verifiably suspended on an informed agent, the public will want to continue and extract yet more. If torture continues on an innocent suspect, the public pressure to suspend torture will be overwhelming. In this way, the two commitment problems that underlie our analysis reappear when torture is verifiable.

As contractual and reputational solutions are problematic, the principal can try to delegate torture to a specialist. In the model, the period-by-period decision whether to continue torture is governed by the principal's perceived cost of torturing c. If the principal is representative of the public at large then c reflects the public's moral objection to torture. Alternatively, c can stand for the opportunity cost of waiting to *begin* torturing the next suspect. While the ultimate performance of the mechanism should be measured by comparing the information revealed with these true costs of torture, it is possible that the overall efficiency can be improved by employing a specialist who perceives a lower cost c'. Such a specialist will be *prepared* to torture more and as a result may be *required* to torture less.

Indeed, a specialist who is a sadist and has a small negative "cost" of torture $c' < 0$, can extract the entire quantity x of information from the informed. A sadist is willing to torture a silent suspect even if there is zero probability he is informed. The informed can give up all his information without compromising the incentive of the specialist to continue to torture a suspect who does not yield anything. It is still the case that in equilib-

[25] See Dewatripont (1989) on contracting, Fudenberg and Tirole (1983), Sobel and Takahashi (1983), Gul, Sonnenschein, and Wilson (1986), and Hart and Tirole (1988) on the renegotiation and the Coase conjecture. See Freixas, Guesnerie, and Tirole (1985) and Laffont and Tirole (1988) on the ratchet effect.

rium the informed suspect must yield a quantity Δ of information units per period. Otherwise, once the suspect has yielded x, the specialist will continue torture for pleasure not for information. The agent can do better by slowing down the release of information and keeping some in hand to buy off the specialist. In this sense, delegation to a specialist with a small benefit to torture can alleviate one of the commitment problems inherent in torture.

But this solution creates other problems. First, there is a difficulty if the specialist is a strong sadist with $\Delta < -c'$ and gets too much enjoyment from torture. A strong sadist has no incentive to demand information and he simply tortures every period. A contractual solution via monetary incentives for the specialist is difficult because torture is unverifiable. The specialist is left to his own devices and a sufficiently strong sadist is impossible to control. Hence, it is important to screen specialists effectively to identify that their incentives are aligned sufficiently with the principal's preferences.

Even is $c' < 0$ is small, the specialist will torture the agent in all periods when he is not extracting information. For example, in the ticking time-bomb scenario, suppose the specialist demands information during the ticking-time bomb phase. He will torture the agent in all the time outside this phase. Hence, an upper bound on the *principal's payoff* is

$$\mu x - \frac{c\,(1-\mu)}{\Delta} - c(T - \frac{x}{\Delta})$$

which is negative when the ticking time-bomb explodes far enough in the future. In the perpetual threat scenario, the situation is even worse with the specialist torturing ad infinitum.

It might seem as if the problem can be resolved by hiring and sacking the specialist at the appropriate time. But this uncovers the deepest problem with the delegation strategy whenever the cost of torture to the specialist differs from the cost to the principal: As torture is unverifiable, the principal can always terminate the specialist at any point in time. In fact, as soon as the agent does not yield information, the principal intervenes, replaces the specialist and stops torture. Then, one of the key commitment problems with torture reappears and our basic analysis is relevant again.

In short, the commitment problems we study are also present in economic environments. They are magnified in the torture environment by the fact that torture is unverifiable.

7 Conclusion

We have made some simplifying assumptions to keep our model tractable and simple. For example, we have only allowed for variable costs to torture but there might be fixed costs. Perhaps there is a psychological cost to even beginning torture. There are additional fixed costs to incarceration in an interrogation facility whether the agent is tortured or not. The marginal decisions of the principal and the agent do not depend on fixed costs and our equilibrium characterization is unchanged. But the principal value of torture is negative as the period length becomes small and hence it is better never to begin. Adding additional elements such as costs of verification reduces the value of torture yet further.

Also, we only allow a high value suspect to have a known quantity of information. Realistically, the quantity of information held by a target may also be unknown. In a natural model, there is positive probability that the agent is uninformed and, if he is informed, his information is drawn from some bounded interval. The value of torture to the principal is lower in this continuum model than in the two type case. This is because the agent captures more information rents. As the principal does not know the quantity of information an informed agent has, he asks for less information than in the two type case after the agent has not confessed. Then, the principal will torture only if the probability that the suspect is informed is high and this in turn implies that the rate of confession must be lower than in the two type case so the principal will torture for fewer periods. All of this means the value of torture is lower in the continuum model. Intuitively, the principal is at an informational disadvantage when he knows less about the information in the hands of the suspect and this can only reduce his welfare.

Our basic message is robust to these variations: The effectiveness of torture as an information extraction mechanism depends crucially on the assumption that it is possible to commit to an incentive scheme. When the principal can revisit his torture strategy at discrete points in time, the informed agent must confess slowly in equilibrium. We show that there is then a maximum amount of time torture will ever be used. This reduces the value of torture and when the principal can revisit the torture decision frequently, the value disappears.

Our main purpose is to study torture as an information extraction mechanism but some of our results apply to other settings. The agent has a privately known wealth level and the principal is collecting taxes. The agent's wealth level is impossible to observe but the principal can inflict some cost - imprisonment or the costs of being audited - on the agent. This kind of intervention is also costly for the principal and he can revisit his policy every period. If the principal finds out the agent is wealthy, he has an incentive to continue to extract more resources. If the principal believes the agent has no wealth, he has an incentive to stop as auditing is costly. Each period the agent can disappear with what remains of his wealth. The principal can either be a legitimate but rapacious government or a criminal organization. Similarly, a central government may have delegated tax collection to an regional authority but does not know if the authority has collected any revenues. In weakly institutionalized environments, the only method of extracting resources from the authority is to threaten a costly conflict if no transfer is forthcoming.[26]

Alternatively, the agent may have resources other than wealth which are valuable both to him and the principal. The agent may hold hostages whom he uses as slaves and the principal's only instrument to persuade the agent to return the hostages is the threat of force. The agent may have nuclear material that is potentially valuable for creating deterrence and the principal can use only mutually costly sanctions or force to extract the material.[27] These applications suggest other extensions - for example tax auditing may release information as well as be costly. These are promising topics for future research.

References

ALEXANDER, M., AND J. R. BRUNING (2008): *How to break a terrorist: the U.S. interrogators who used brains, not brutality, to take down the deadliest man in Iraq*. Free Press, New York, 1st free press hardcover ed edn.

[26] See Andreoni, Erard, and Feinstein (1998) for a survey of tax compliance and Polinsky and Shavell (1984) for a model of imprisonment with unobservable wealth when the principal can fully commit. (Schelling, 1984, Chapter 8) argues that the mafia will find it difficult to extort a criminal whose income is unobservable.

[27] Thompson and Heather (1996) describes the extortion of tribute and fugitives by Atilla the Hun from the Romans. The Romans would sometimes deny fugitives valued by Attila were in their possession. The Huns would be forced to attack to extract them.

ANDREONI, J., B. ERARD, AND J. FEINSTEIN (1998): "Tax compliance," *Journal of economic literature*, pp. 818–860.

AUGUSTINE, S. (1825): *De civitate Dei: libri XXII.*, vol. 2. Caroli Tauchnitii.

AUSUBEL, L. M., AND R. J. DENECKERE (1992): "Bargaining and the right to remain silent," *Econometrica: Journal of the Econometric Society*, pp. 597–625.

DERSHOWITZ, A. M. (2002): *Why terrorism works: Understanding the threat, responding to the challenge*. Yale University Press.

DEWATRIPONT, M. (1989): "Renegotiation and information revelation over time: the case of optimal labor contracts," *The Quarterly Journal of Economics*, 104(3), 589–619.

FAY, G. R. (2004): "AR 15-6 investigation of the Abu Ghraib detention facility and 205th Military Intelligence Brigade," *United States Department of Defense Detainees Investigations*.

FREIXAS, X., R. GUESNERIE, AND J. TIROLE (1985): "Planning under incomplete information and the ratchet effect," *The Review of Economic Studies*, 52(2), 173–191.

FUCHS, W., AND A. SKRZYPACZ (2011): "Bargaining with Deadlines and Private Information," Discussion paper, Mimeo, University of California Berkley.

FUDENBERG, D., AND D. LEVINE (1989): "Reputation and equilibrium selection in games with a patient player," *Econometrica: Journal of the Econometric Society*, 57(4), 759–778.

——— (1992): "Maintaining a reputation when strategies are imperfectly observed," *The Review of Economic Studies*, pp. 561–579.

FUDENBERG, D., AND J. TIROLE (1983): "Sequential bargaining with incomplete information," *The Review of Economic Studies*, 50(2), 221–247.

GUL, F., H. SONNENSCHEIN, AND R. WILSON (1986): "Foundation of dynamic monopoly and the coase conjecture," *Journal of Economic Theory*.

Hart, O. (1989): "Bargaining and strikes," *The Quarterly Journal of Economics*, 104(1), 25–43.

Hart, O., and J. Tirole (1988): "Contract renegotiation and Coasian dynamics," *The Review of Economic Studies*, 55(4), 509–540.

Horner, J., and L. Samuelson (2009): "Managing Strategic Buyers," http://pantheon.yale.edu/ ls529/papers/MonoPrice10.pdf.

Horner, J., and A. Skrzypacz (2015): "Selling information," *Journal of Political Economy*.

Kreps, D., and R. Wilson (1982): "Reputation and imperfect information," *Journal of economic theory*, 27(2), 253–279.

Laffont, J., and J. Tirole (1988): "The dynamics of incentive contracts," *Econometrica*, 56(5), 1153–1175.

Leibovich, M. (2009): "New Cheney Taking Stage for the GOP," *New York Times*.

Mialon, H., S. Mialon, and M. Stinchcombe (2012): "Torture in Counterterrorism: Agency Incentives and Slippery Slopes," *Journal of Public Economics*, 96(1-2), 33–41.

Myerson, R. B. (1991): *Game theory: analysis of conflict.* Cambridge. Harvard University.

Padro i Miquel, G., and P. Yared (2010): "The Political Economy of Indirect Control," *Quarterly Journal of Economics*, 127, 947–1015.

Polinsky, A. M., and S. Shavell (1984): "The optimal use of fines and imprisonment," *Journal of Public Economics*, 24(1), 89–99.

Post, J. M. (2005): *Military studies in the Jihad against the tyrants: the Al-Qaeda training manual.* USAF Counterproliferation Center, Maxwell Air Force Base, Ala.

Schelling, T. C. (1984): *Choice and consequence.* Harvard University Press.

SOBEL, J., AND I. TAKAHASHI (1983): “A multistage model of bargaining,” *The Review of Economic Studies*, 50(3), 411–426.

THOMPSON, E. A., AND P. J. HEATHER (1996): *The Huns*. Blackwell Oxford.

WALZER, M. (1973): “Political action: The problem of dirty hands,” *Philosophy & public affairs*, 2(2), 160–180.

WILKERSON, L. (2009): “The Truth About Richard Bruce Cheney,” *Washington Note*.

A Proofs for the Perpetual Threat Scenario

Proof of Proposition 1. Consider the following strategies.

- As long as the suspect has not yet confessed the principal demands $y = 0$.
- If the suspect has previously conceded $z > 0$ then the principal demands $y = \max\{0, x - z\}$.
- If the suspect has not yet confessed then he refuses any demand.
- If the suspect has previously conceded $z > 0$ then the suspect agrees to any demand no larger than $\max\{0, x - z\}$.

Given the principal’s strategy, the suspect knows that any concession will lead him to concede all of x within the next period. Thus his payoff from confessing is no larger than $-\delta x$. Since $\Delta < \delta x$ the suspect prefers to withstand a single period of torture (after which the principal will no longer torture, as specified by the strategy above.)

If the suspect has already confessed he knows that the principal will continue to demand the remaining information until it is forthcoming. The suspect’s strategy of conceding is therefore optimal (rather than withstand torture before eventually conceding anyway.)

Given that the suspect will refuse any demand the principal optimally does not torture. And if the suspect has previously conceded, because the suspect's strategy specifies further concessions the principal optimally demands the remainder of the information.

We have shown that the strategies are sequentially rational and therefore an equilibrium.

□

Proof of Lemma 1. Let $\epsilon > 0$ satisfy

$$\epsilon x - (1 - \epsilon)c = 0.$$

Along a history in which the suspect has not confessed and torture continues, the posterior probability that he is informed is declining monotonically. It therefore converges to some limit. In particular there is an integer n such that after n periods of torture the posterior μ is within ϵ of its limit. Let q be the total probability that the informed suspect confesses throughout the remainder of the game. Then

$$\mu - \frac{(1-q)\mu}{1-q\mu} < \epsilon$$

because the second term on the left-hand side is the conditional probability that a resistant suspect is informed if the total concession probability is q. Simplifying this inequality gives

$$\mu q < \epsilon,$$

and therefore, following the nth period of torture, the principal's continuation value from carrying on torturing is at most

$$\mu q x - (1 - \mu q)c$$

which is negative. The unique sequentially rational continuation strategy is therefore to halt torture after torturing for n periods. Since any equilibrium strategy for the principal is a probability distribution over sequentially rational pure strategies, with probability 1 torture cannot last more than n periods. □

The following lemma states that if the suspect is induced to concede in some way that leads to updated posterior $\tilde{\mu}_k$, then the total concession probability is the same whether these concessions happen over the course of two periods or in a single period. This is then used in Lemma 4 below to show that the principal is better off the faster the posterior reaches $\tilde{\mu}_k$, i.e the principal prefers to frontload concessions.

Lemma 3. *For any $\mu \in (0,1)$ and $q \in (0,1)$,*

$$q + (1-q)\tilde{q}_k(B(q;\mu)) = \tilde{q}_k(\mu). \tag{14}$$

Proof. The equality follows immediately from the fact that $B(\cdot;\mu)$ applied to either side yields $\tilde{\mu}_{k-1}$ and that $B(q;\mu)$ is invertible. For the right-hand side this is by definition. The calculation for the left-hand side follows.

$$\begin{aligned} B(q + (1-q)\tilde{q}_k(B(q;\mu));\mu) &= \frac{\mu\left(1 - [q + (1-q)\tilde{q}_k(B(q;\mu))]\right)}{1 - \mu\left[q + (1-q)\tilde{q}_k(B(q;\mu))\right]} \\ &= \frac{\frac{\mu(1-q)}{1-\mu q}\left[1 - \tilde{q}_k(B(q;\mu))\right]}{1 - \frac{\mu(1-q)\tilde{q}_k(B(q;\mu))}{1-\mu q}} \\ &= \frac{B(q;\mu)\left[1 - \tilde{q}_k(B(q;\mu))\right]}{1 - B(q;\mu)\tilde{q}_k(B(q;\mu))} \\ &= B(\tilde{q}_k(B(q;\mu)); B(q;\mu)) \\ &= \tilde{\mu}_{k-1}. \end{aligned}$$

□

The intuition for the following lemma is that the principal prefers to frontload concessions. The principal benefits from earlier concessions for two reasons. First the total cost of torture will be reduced and second, the principal will have more time to extract additional information from a suspect who concedes earlier.

Lemma 4. *For all k and for any μ, the expression*

$$k\Delta\mu q + (1-\mu q)\left[\tilde{V}^{k-1}(B(q;\mu)) - c\right] \tag{15}$$

is strictly increasing in q.

Proof. Define $Z(q) = B(q;\mu)\tilde{q}_{k-1}(B(q;\mu))$, and substitute into the definition of $\tilde{V}^{k-1}(B(q;\mu))$:

$$\tilde{V}^{k-1}(B(q;\mu)) = Z(q)(k-1)\Delta - c(1 - Z(q)).$$

Substituting into Equation 15, we have

$$k\Delta\mu q + (1-\mu q)\left[Z(q)(k-1)\Delta - c(1-Z(q)) - c\right].$$

This can be re-arranged as follows.

$$\mu q\left[\Delta + c\right] + \left[\mu q + (1-\mu q)Z(q)\right]\left[(k-1)\Delta + c\right] - 2c. \tag{16}$$

We will prove that the second term is constant in q and therefore that the overall expression is strictly increasing in q. By Lemma 3,

$$q + (1-q)\tilde{q}_{k-1}(B(q;\mu)) = \tilde{q}_{k-1}(\mu)$$

If we multiply both sides by μ

$$\mu q + \mu(1-q)\tilde{q}_{k-1}(B(q;\mu)) = \mu\tilde{q}_{k-1}(\mu)$$

and then multiply the second term on the left-hand side by 1,

$$\mu q + \frac{\mu(1-q)\tilde{q}_{k-1}(B(q;\mu))(1-\mu q)}{(1-\mu q)} = \mu\tilde{q}_{k-1}(\mu)$$

we obtain

$$\mu q + (1-\mu q)B(q;\mu)\tilde{q}_{k-1}(B(q;\mu)) = \mu\tilde{q}_{k-1}(\mu)$$

or

$$\mu q + (1-\mu q)Z(q) = \mu\tilde{q}_{k-1}(\mu)$$

establishing that the second term in Equation 16 is constant in q. □

Proof of Lemma 2. Take any equilbrium and let n be the maximum number of periods of torture among all pure strategies in the support of the principal's mixed strategy. Such an n exists by Lemma 1. Take any pure strategy in the support of the principal's mixed strategy which tortures for n periods. We will establish a bound on the payoff to this pure strategy. Since all strategies in the support of the principal's equilibrium strategy yield the same payoff this will deliver the result.

Let μ_1 be the posterior entering the last period of torture. The principal's continuation payoff entering the last period of torture is at most

$$\mu_1\Delta - (1-\mu_1)c.$$

To see why, note that the suspect can secure a payoff of $-\Delta$ by resisting torture one last time. Therefore the suspect's payoff from confessing can be no less than $-\Delta$ implying that he concedes at most Δ to the principal. Note that this bound equals $\tilde{V}^1(\mu_1)$.

Now suppose that for any $\tilde{V}^{k-1}(\mu_{k-1})$ gives an upper bound on the principal's payoff when there are $k-1$ periods of torture remaining and the suspect is informed with probability μ_{k-1}. Let μ_k be the probability the suspect is informed entering the kth-to-last period of torture.[28] Let q be the probability with which the informed suspect confesses in that period. The principal's continuation payoff entering that period is bounded by

$$\mu_k q k\Delta + (1-\mu_k q)\left[\tilde{V}^{k-1}(B(q;\mu_k)) - c\right].$$

To see why note that the suspect can secure a payoff of at least $-k\Delta$ from resisting torture for the k remaining periods.[29] Thus the principal can extract at most $k\Delta$ from a suspect who concedes in the kth-to-last period of torture. In the event that the agent does not concede, the principal incurs the cost c and later obtains the continuation value from resuming torture which we have already bounded by $\tilde{V}^{k-1}(\cdot)$. As this payoff comes later, it would be discounted but that only lowers the principal's payoff even further.

Because the principal's strategy prescribes an additional $k-1$ periods of torture later if the suspect does not concede, and the principal's strategy is sequentially rational, the continuation value to the principal from doing so must be non-negative, i.e. $\tilde{V}^{k-1}(B(q;\mu_k)) \geq 0$. Thus, the following constrained maximization represents an upper bound on the principal's payoff entering the kth-to-last period in which he tortures.

$$\max_q \mu_k q k\Delta + (1-\mu_k q)\left[\tilde{V}^{k-1}(B(q;\mu_k)) - c\right] \tag{17}$$

$$\text{such that } \tilde{V}^{k-1}(B(q;\mu_k)) \geq 0 \tag{18}$$

We have shown in Lemma 4 that the maximand is strictly increasing in q. Since moreover $\tilde{V}^{k-1}(B(q;\mu_k))$ is strictly decreasing in q, the

[28] Recall that we are considering a pure strategy for the principal so there is a well-defined subsequence of periods in which he makes a non-trivial demand and threatens torture.

[29] Note that the costs from subsequent periods of torture would be discounted by the suspect. Thus the suspsect's payoff is *strictly* larger than $-k\Delta$)

constraint binds. Thus the maximum is achieved by the q that satisfies $\tilde{V}^{k-1}(B(q,\mu_k)) = 0$, i.e. $\tilde{q}_k(\mu_k)$ and thus the principal's payoff is bounded by

$$\mu_k \tilde{q}_2(\mu_k) k\Delta - (1 - \mu_k \tilde{q}_k(\mu_k))c,$$

which is simply $\tilde{V}^k(\mu_k)$. □

Proof of Theorem 2. If the principal begins torturing in period k, then his payoff must be non-negative. By Lemma 2 $\tilde{V}^k(\mu_0)$ is an upper bound for the principal's payoff and hence $\tilde{V}^k(\mu_0) \geq 0$. In particular $\mu_0 \geq \tilde{\mu}_k$. Since $\tilde{\mu}_j \geq \tilde{\mu}_{j-1}$ for all j, we have $\mu_0 \geq \tilde{\mu}_j$ for all $j = 1, \ldots k$. By the definition of $\tilde{V}^j(\tilde{\mu}_j)$,

$$0 = \tilde{V}^j(\tilde{\mu}_j) = \tilde{\mu}_j \tilde{q}_j(\tilde{\mu}_j) j\Delta - c(1 - \tilde{\mu}_j \tilde{q}_j(\tilde{\mu}_j))$$

Re-arranging and using the definition of $\tilde{q}_j(\mu_j)$,

$$\frac{\tilde{\mu}_j - \tilde{\mu}_{j-1}}{1 - \tilde{\mu}_{j-1}} = \tilde{\mu}_j \tilde{q}_j(\tilde{\mu}_j) = \frac{c}{j\Delta + c}$$

Since $\tilde{\mu}_j \leq \mu_0$ for all $j = 1, \ldots, k$,

$$\tilde{\mu}_j - \tilde{\mu}_{j-1} \geq (1 - \mu_0) \left[\frac{c}{j\Delta + c} \right]$$

Thus,

$$\mu_0 \geq \tilde{\mu}_k \geq \sum_{j=1}^{k} (1 - \mu_0) \left[\frac{c}{j\Delta + c} \right]$$

and therefore $k \leq K(\mu_0)$, establishing the first part of the theorem. The second part then follows from Lemma 2. The third part is a crude bound that calculates only the maximum amount of information that can be extracted from the informed in $K(\mu_0)$ periods. □

B Proofs for the Ticking Time-Bomb Scenario

Proof of Proposition 2. First suppose that $k = 1$ so that there is a single period remaining and assume that the suspect has revealed all but the quantity $\tilde{x}$ of information. Suppose that he is asked to reveal $y \leq \tilde{x}$ or else endure torture. Since there is a single period remaining, the principal is threatening to inflict Δ on the suspect. If $y > \Delta$ the suspect will refuse, if $y < \Delta$, the suspect strictly prefers to reveal y and if $y = \Delta$ he is indifferent. The unique equilibrium is for the principal to ask for $y = \min\{\tilde{x}, \Delta\}$ and for the suspect to reveal y. This gives the suspect a payoff of $-\min\{\tilde{x}, \Delta\}$. Now to prove the lemma by induction, suppose that in all equilibria, the complete information continuation game beginning in period $k-1$ with $\tilde{x}$ yet to be revealed yields the payoff

$$-\min\{\tilde{x}, (k-1)\Delta\}$$

to the suspect and $\min\{\tilde{x}, (k-1)\Delta\}$ for the principal and assume that there are k periods remaining and $\tilde{x}$ has yet to be revealed. Suppose the suspect is asked in period k to reveal $y \leq \min\{\tilde{x}, \Delta\}$ or else endure torture. If the suspect complies he obtains payoff

$$-\left[y + \min\left\{\tilde{x} - y, (k-1)\Delta\right\}\right]$$

and if he refuses his payoff is

$$-\left[\Delta + \min\left\{\tilde{x}, (k-1)\Delta\right\}\right]$$

which is weakly smaller and strictly so when $y < \Delta$. So the suspect will strictly prefer to reveal if $y < \Delta$ and he will be indifferent when $y = \Delta$. It follows that for any $\varepsilon > 0$, if the principal asks for $\min\{\tilde{x}, \Delta\} - \varepsilon$, sequential rationality requires that the suspect complies. By the induction hypothesis this leads to a total payoff of $\min\{\tilde{x}, k\Delta\} - \varepsilon$ for the principal. Since $\min\{\tilde{x}, k\Delta\}$ is the maximum payoff for the principal consistent with feasibility and individual rationality for the suspect, it follows that all equilibria must yield $\min\{\tilde{x}, k\Delta\}$ for the principal.[30] Any strategy profile which gives this payoff to the principal must involve maximal revelation ($\min\{\tilde{x}, k\Delta\}$) and no torture. Thus, all equilibria give payoff $-\min\{\tilde{x}, k\Delta\}$ to the suspect. □

Proof of Proposition 3. Consider any period $k+1$ within the ticking time-bomb phase when the suspect has yet to confess and the principal demands $y > 0$. By Proposition 2 the payoff to a suspect confesses is $-y - k\Delta$. Suppose there is a later period such that if the suspect has not yet

[30] In fact if $k\Delta > \tilde{x}$ then there are multiple equilibria all yielding this payoff, corresponding to various sequences of demands adding up to $\tilde{x}$.

confessed the principal does not torture. Then the payoff to a suspect who stays silent in periods $k+1$ until the end is $-k\Delta$. That is the suspect strictly prefers to stay silent contradicting the assumption that there was effective torture in period $k+1$. □

Proof of Proposition 4. Suppose that there are two periods of effective torture prior to the ticking time-bomb phase, periods k and j with $k > j$. Since there is effective torture, the suspect confesses with positive probability in each period. For that to be sequentially rational for the suspect the payoff from confessing must be no smaller than the payoff to resisting. By Proposition 2, in both periods j and k the payoff to confessing is $-x$. Thus, the payoff to resisting is smaller than or equal to $-x$ in period j. The payoff to resisting in period k is equal to the cost of torture in period k, i.e. $-\Delta$ plus the continuation payoff. The continuation payoff is no larger than the continuation payoff in period j from either confessing or resisting. Each of those payoffs are less than or equal to $-x$. Thus, the payoff to a suspect who resists in period k is less than or equal to $-\Delta - x$. Since the suspect can guarantee a payoff of $-x$ by confessing, this implies that the informed suspect confesses with probability 1 in period k. But then a suspect who resists in period k is certain to be uninformed and the only sequentially rational strategy for the principal is to stop torturing, contradicting Proposition 3. □

Lemma 5. *The system of equations Equation 9-Equation 11 uniquely defines for each $k = 2, \ldots \bar{k}+1$ the value μ^*_k, and the functions $q_k(\cdot)$ and $V^k(\cdot)$ over the range $[\mu^*_{k-1}, 1]$. The functions $V^k(\cdot)$ are linear in μ with slopes increasing in k, and $V^k(\mu^*_k) > 0$ for all $k = 2, \ldots, \bar{k}+1$*

Proof. By Equation 3 and Equation 11,

$$\mu q_k(\mu) = \frac{\mu - \mu^*_{k-1}}{1 - \mu^*_{k-1}}$$

and hence we can write $V^k(\mu)$ as follows

$$V^k(\mu) = \frac{\mu - \mu^*_{k-1}}{1 - \mu^*_{k-1}} \left(\min\{x, k\Delta\} + c - V^{k-1}(\mu^*_{k-1}) \right) + V^{k-1}(\mu^*_{k-1}) - c$$

showing that $V^k(\cdot)$ is linear in μ. Evaluating at $\mu = \mu^*_{k-1}$ and $\mu = 1$, we see that

$$V^k(\mu^*_{k-1}) < V^{k-1}(\mu^*_{k-1}) \qquad V^k(1) \geq V^{k-1}(1)$$

and therefore the value μ_k^* defined in Equation 10 is unique. This in turn implies that the functions $q_{k+1}(\cdot)$ and $V^{k+1}(\cdot)$ are uniquely defined. □

B.1 Full Description and Verification of Equilibrium

In period k^* the principal begins torturing with probability 1 and making the demand $y = \Delta$. The informed agent yields Δ with probability less than 1, after which he subsequently reveals an additional Δ in each of the remaining periods until either the game ends or he reveals all of x. With the complementary probability, he remains silent. As long as the agent has remained silent, in particular if he is uninformed, the torture continues with demands of Δ until the end of the game. The principal demands Δ with probability 1 in periods $k < \bar{k}$ and with a probability less than one in period $\bar{k}$ (if $k^* = \bar{k} + 1$.)

First, since the informed agent concedes in period k^* with probability $q_{k^*}(\mu_0)$, the posterior probability that he is informed after he resists in period k^* is $\mu_{k^*-1}^*$ by Equation 11. In all periods $1 < k < k^*$, if he has yet to concede, he makes his first concession with probability $q_k(\mu_k^*)$. Hence again by Equation 11, the posterior will be μ_k^* at the beginning of any period $k < k^* - 1$ in which he has resisted in all periods previously.

In period 1, if the suspect has yet to concede the principal tortures with probability 1 and the informed agent yields with probability 1. If μ is the probability that the agent is informed, the principal obtains payoff Δ with probability μ and incurs cost c with probability $1 - \mu$. Thus the principal's payoff in period 1, the final period, is

$$V^1(\mu) = \Delta\mu - c(1 - \mu).$$

Since in equilibrium the posterior probability will be μ_1^*, the principal's payoff continuation payoff is $V^1(\mu_1^*)$ which is zero by the definition of μ_1^*.

By induction, the principal's continuation payoff in any period $k \leq k^*$ in which the agent has yet to concede is given by

$$V^k(\mu) = \mu q_k(\mu) \min\{x, k\Delta\} + (1 - \mu q_k(\mu)) \left[V^{k-1}(\mu_{k-1}^*) - c \right]$$

if the posterior probability that the agent is informed is μ. This is because the informed agent concedes with probability $q_k(\mu)$ and subsequently gives Δ in all remaining periods until x is exhausted. In the event the agent does not concede, the principal incurs cost c and obtains the continuation value $V^{k-1}(\mu^*_{k-1})$. In equilibrium in period k the probability that the agent is informed conditional on previous resistance is μ^*_k for $k < k^*$ and μ_0 in period k^*. Since prior to period k^*, the principal obtains no information and incurs no cost of torture, his equilibrium payoff is $V^{k^*}(\mu_0)$, and his continuation payoff after resistance up to period $k < k^*$ is $V^k(\mu^*_k)$.

When the suspect resists torture prior to period k and the posterior is μ^*_k, by definition $V^k(\mu^*_k) = V^{k-1}(\mu^*_k)$. This means that the principal is indifferent between his equilibrium continuation payoff $V^k(\mu^*_k)$, and the payoff he would obtain if he were to "pause" torture for one period (set $y = 0$) and resume in period $k-1$. Moreover, by Lemma 5, this payoff is strictly higher than waiting for more than one period (this is illustrated in Figure 1.) Thus the principal's strategy to demand $y = \Delta$ with probability 1 in periods $1, \ldots, \bar{k}-1$ is sequentially rational.

When the suspect has revealed himself to be informed, the principal in equilibrium extracts the maximum amount of information $k\Delta$ given the remaining periods.

Turning to the suspect, in periods $1, \ldots \bar{k}$, his continuation payoff is $-k\Delta$ whether he resists torture or concedes. This is because by conceding he will eventually yield a total of $k\Delta$, and by resisting he will be tortured for k periods which has cost $k\Delta$. His strategy of randomizing is therefore sequentially rational in these periods. [31]

Next we describe the behavior after a deviation from the path. If the suspect has revealed information previously then he accepts any demand for information less than or equal to the amount he would eventually be revealing in equilibrium. That is, if there are k periods remaining and z is the quantity of information yet to be revealed, he will accept a demand to reveal y if and only if $y \leq \min\{z, k\Delta\}$. The principal ignores any deviations by the suspect along histories where the suspect has already revealed information. If no information has been revealed yet, then behavior after

[31] Period $\bar{k}+1$ is a special case. In this period yielding will give the suspect a payoff of $-x$ (the time constraint is not binding). If instead he resists, his payoff is

$$-\Delta - \rho\bar{k}\Delta - (1-\rho)(\bar{k}-1)\Delta$$

because the principal randomizes between continuing torture in the following period and waiting for one period before continuing. By the definition of ρ this payoff equals x and so the suspect is again indifferent and willing to randomize.

a deviation by the principal depends on whether $k^* < \bar{k}+1$ or $k^* = \bar{k}+1$ and on the value of the current posterior probability μ that the suspect is informed. (Note that this posterior is always given by Bayes' rule because the presence of an uninformed type means that no revelation is always on the path.) First consider the case $k^* < \bar{k}+1$. Suppose $k \leq k^*+1$ then the suspect refuses any demand y greater than Δ. On the other hand if the principal deviates and asks for $0 < y \leq \Delta$, then the suspect concedes with the equilibrium probability $q_k(\mu)$. To maintain incentives the principal must then alter his continuation strategy (unless $k = 1$ in which case the game ends.) In particular, after deviating and demanding $0 < y < \Delta$, if the suspect resists, then in period $k-1$, the principal will randomize with the probability $\rho(y) = \rho/\Delta$ that ensures that the agent was indifferent in period k between conceding (eventually yielding $y + (k-1)\Delta$) and resisting:

$$y + (k-1)\Delta = \Delta + \rho(y)\Delta + (k-2)\Delta.$$

If instead $k > k^*+1$ then the suspect refuses any demand and the principal reverts to the equilibrium continuation and waits to resume torture in period k^*. Next suppose $k^* = \bar{k}+1$. If $k \leq \bar{k}+1$ then deviations by the principal lead to identical responses as in the previous case of $k \leq k^*+1$ when $k^* < \bar{k}+1$. The last subcase to consider is $k > \bar{k}+1$. If $y > x$ then the suspect refuses with probability 1. If $y \leq x$ then the deviation alters the continuation strategies in two ways. First, the informed suspect yields to the demand with probability $q_{\bar{k}+1}(\mu)$. If he does concede, he will ultimately yield all of x because there will be at least $\bar{k}+1$ additional periods of torture to follow. Second, the principal subsequently pauses torture until period $\bar{k}$ at which point he begins torturing with probability ρ. Effectively, this deviation has just shifted the torture that would have occurred in period $\bar{k}+1$ to the earlier period k.

B.2 Proof of Theorem 3

Proof of Theorem 3. Because Proposition 2 characterizes continuation equilibria following a concession, the analysis focuses on continuation equilibria following histories in which the suspect has yet to concede, and the posterior probability of an informed suspect is μ. So when we say that "there is torture in period k" we mean that upon reaching period k without a concession, principal demands $y > 0$.

We first consider continuation equilibria starting in a period $k \leq \bar{k}$ in which there is torture in period k. We show by induction on $k = 1, \ldots, \bar{k}$ that if there is torture in period k, then the principal's continuation equilibrium payoff beginning from period k is $V^k(\mu)$. We begin with the case of $k = 1$. Suppose that the game reaches period 1 with no concession and a posterior probability μ that the suspect is informed. In this case the continuation equilibrium is unique. Indeed, any demand $y < \Delta$ will be accepted by the informed and any demand $y > \Delta$ would be rejected. If the principal makes any positive demand he will therefore demand $y = \Delta$ and the informed agent will concede. This yields the payoff $\mu\Delta - (1-\mu)c$. In particular, when $\mu > \mu_1^*$, the unique equilibrium is for the principal to demand $y = \Delta$ and when $\mu < \mu_1^*$ the principal demands $y = 0$. In the former case the agent's payoff is $-\Delta$ and in the latter zero. In the case of $\mu = \mu_1^*$ there are multiple equilibria which give the principal a zero payoff and the agent any payoff in $[0, -\Delta]$.

Next, as an inductive hypothesis, we assume the following is true of any continuation equilibrium beginning in period $k-1 < \bar{k}$ with posterior μ.

1. If $\mu > \mu_{k-1}^*$and there is torture with positive probability in period $k-1$ then the principal's payoff is $V^{k-1}(\mu)$ and the agent's payoff is $-(k-1)\Delta$.

2. If $\mu = \mu_{k-1}^*$ and there is torture with positive probability in period $k-1$ then the principal's payoff is $V^{k-1}(\mu)$ and the agent's payoff is any element of $[-(k-2)\Delta, (-k-1)\Delta]$.

3. If $\mu < \mu_{k-1}^*$ then there is no continuation equilibrium with torture with positive probability in period $k-1$.

Now, consider any continuation equilibrium beginning in period k with a positive demand $y > 0$. First, it follows from Proposition 2 that $y \leq \Delta$. For if the informed suspect yields $y > \Delta$ in period $k \leq \bar{k}$ his payoff would be smaller than $-k\Delta$ which is the least his payoff would be if he were to resist torture for the rest of the game. The suspect will therefore refuse any demand $y > \Delta$ and such a demand would yield no information and no change in the posterior probability that the agent is informed. Because torture is costly and the induction hypothesis implies that the principal's payoff is determined by the posterior, the principal would strictly prefer $y = 0$ in period k, a contradiction.

Assume that the informed concedes with probability q. If $q > q_k(\mu)$ then $B(q;\mu) < \mu^*_{k-1}$ and the induction hypothesis, there will be no torture in period $k-1$ if the suspect resists in period k. This means that a resistant suspect has a payoff no less than $-(k-1)\Delta$. But if the suspect concedes in period k, by Proposition 2, his payoff will be $-y-(k-1)\Delta$. The informed suspect cannot weakly prefer to concede, a contradiction.

Thus, $q \leq q_k(\mu)$. Now suppose $y < \Delta$. In this case we will show that $q \geq q_k(\mu)$ so that $q = q_k(\mu)$. For if $q < q_k(\mu)$, i.e. $B(q;\mu) > \mu^*_{k-1}$ then by the induction hypothesis the continuation equilibrium after the suspect resists gives the suspect a payoff of $-(k-1)\Delta$ for a total of $-k\Delta$. But conceding gives $-y-(k-1)\Delta$ by Proposition 2 and thus the suspect strictly prefers to concede, a contradiction since $q < q_k(\mu)$ requires that the suspect weakly prefers to resist.

We have shown that if $y < \Delta$ then the informed suspect concedes with probability $q_k(\mu)$. This yields payoff to the principal

$$W(y) = \mu q_k(\mu)\left[y + (k - 1\Delta)\right] + (1 - \mu q_k(\mu))\left[V^{k-1}(\mu^*_{k-1}) - c\right]$$

because a conceding suspect will subsequently give up $(k-1)\Delta$, because $B(q_k(\mu);\mu) = \mu^*_{k-1}$, and because the induction hypothesis implies that the principal's continuation value is given by V^{k-1}.

Since this is true for all $y > 0$ and in equilibrium the principal chooses y to to maximize his payoff, it follows that the principal's equilibrium payoff is at least

$$\sup_{y<\Delta} W(y) = W(\Delta) = V^k(\mu).$$

Moreover, since $W(y)$ is strictly increasing in y, it follows that the principal must demand $y = \Delta$. We have already shown that the informed suspect concedes with a probability no larger than $q_k(\mu)$. We conclude the inductive step by showing that he concedes with probability equal to $q_k(\mu)$ (this was shown previously only under the assumption that $y < \Delta$) and therefore that the principal's payoff is exactly $V^k(\mu)$.

Suppose that the informed suspect concedes with a probability $q < q_k(\mu)$. Then, conditional on the suspect resisting, the posterior probability he is informed will be $B(q;\mu) < \mu^*_{k-1}$. By the induction hypothesis, the principal's continuation payoff is $V^{k-1}(B(q;\mu))$ and his total payoff is

$$k\Delta\mu q + (1-\mu q)\left[V^{k-1}(B(q;\mu)) - c\right] \tag{19}$$

(applying Proposition 2.) Note that this equals $V^k(\mu)$ when $q = q_k(\mu)$. We will show that the expression is strictly increasing in q. Since the principal's payoff is at least $V^k(\mu)$, it will follow that the suspect must concede with probability $q_k(\mu)$.

Let us write $Z(q) = B(q;\mu)q_{k-1}(B(q;\mu))$, and with this notation write out the expression for $V^{k-1}(B(q;\mu))$.

$$V^{k-1}(B(q;\mu)) = (k-1)\Delta Z(q) + (1-Z(q))\left[V^{k-2}(\mu^*_{k-2}) - c\right].$$

Substituting into Equation 19, we have the following expression for the principal's payoff.

$$k\Delta\mu q + (1-\mu q)\left[(k-1)\Delta Z(q) + (1-Z(q))\left[V^{k-2}(\mu^*_{k-2}) - c\right] - c\right]$$

This can be re-arranged as follows.

$$\begin{aligned}\mu q\,[\Delta + c] \\ &+ [\mu q + (1-\mu q)Z(q)]\left[(k-1)\Delta + c + V^{k-2}(\mu^*_{k-2})\right] \\ &\qquad + V^{k-2}(\mu^*_{k-2}) - 2c\end{aligned} \tag{20}$$

and now following the same derivation (applying Lemma 3 and manipulating) as in the proof of Lemma 4, we can prove that the second term is constant in q and therefore the overall expression is indeed strictly increasing in q.

We have shown that if there is torture with positive probability in period k then the principal's payoff is $V^k(\mu)$. If $\mu > \mu^*_k$ then $V^k(\mu) > V^l(\mu)$ for all $l < k$ and therefore the principal strictly prefers to begin torture in period k than to wait until any later period. Hence the suspect faces torture for k periods and his payoff is $-k\Delta$. If $\mu = \mu^*_k$ then $V^k(\mu) = V^{k-1}(\mu)$ and the principal can randomize between beginning torture in period k and waiting for one period. The suspect's payoff is therefore any element of $[-(k-1)\Delta, -k\Delta]$. Finally if $\mu < \mu^*_k$, then $V^k(\mu) < V^{k-1}(\mu)$ and the principal strictly prefers to delay the start of torture for (at least) 1 period. Hence in this case the probability of torture in period k is zero. These conclusions establish the inductive claims and conclude the first part of the proof.

To complete the proof, note that we have shown that any equilibrium that commences torture in period $j \leq \bar{k}$ has payoff $V^j(\mu_0)$. It follows from Proposition 4 that any equilibrium that commences torture in period $j > \bar{k}$ has payoff $V^{\bar{k}+1}(\mu_0)$. Since the principal can demand $y = 0$ until the period k that maximizes this payoff function, his equilibrium payoff must be $\max_{k \leq \bar{k}+1} V^k(\mu_0)$. □

C Proofs for Section 3.3

Proof of Theorem 4. The proof is by induction on k. First, the claim holds by definition for $k = 1$. For $k = 2$, note that $\mu_1^* = \tilde{\mu}_1$ and $V^1(\mu_1^*) = 0$, so that $q_2(\cdot) = \tilde{q}_2(\cdot)$ and $V^2(\cdot) \equiv \tilde{V}^2(\cdot)$. Now assume that $\tilde{V}^{k-1} \geq V^{k-1}$. Since the principal's continuation payoff must be non-negative and the functions V^k and $\tilde{V}^k$ are strictly increasing,

$$0 \leq V^{k-2}(\mu_{k-2}^*) < V^{k-2}(\mu_2^*) = V^{k-1}(\mu_{k-1}^*) \leq \tilde{V}^{k-1}(\mu_{k-1}^*).$$

which by the definition of $\tilde{\mu}_{k-1}$ implies $\mu_{k-1}^* > \tilde{\mu}_{k-1}$. This yields the first conclusion $\tilde{q}_k(\cdot) > q_k(\cdot)$. By the definition of V^k,

$$V^k(\mu) = \mu q_k(\mu) \min\{x, k\Delta\} + (1 - \mu q_k(\mu)) \left[V^{k-1}(\mu_{k-1}^*) - c \right]$$

which by the induction hypothesis is bounded by

$$V^k(\mu) \leq \max_{q \leq \tilde{q}_k(\mu)} \left\{ \mu q k\Delta + (1 - \mu q) \left[\tilde{V}^{k-1}(B(q; \mu)) - c \right] \right\}$$

since $q_k(\mu)$ satisfies the constraint and $\mu_{k-1}^* = B(q_k(\mu); \mu)$.

By Lemma 4 the maximand is strictly increasing in q and therefore since $q_k(\mu) < \tilde{q}_k(\mu)$ we have

$$V^k(\mu) < \mu \tilde{q}_k(\mu) k\Delta + (1 - \mu \tilde{q}_k(\mu)) \left[\tilde{V}^{k-1}(B(\tilde{q}_k(\mu); \mu)) - c \right]$$

and since $(B(\tilde{q}_k(\mu); \mu)) = \tilde{\mu}_{k-1}$ we have $\tilde{V}^{k-1}(B(\tilde{q}_k(\mu); \mu)) = 0$ and the right-hand side equals $\tilde{V}^k(\mu)$.

To prove the last claim, we can use the result that $V^k(\mu) \leq \tilde{V}^k(\mu)$ and then proceed through identical steps as in the proof of Theorem 2. □