Torture*

Sandeep Baliga [†] Jeffrey C. Ely [‡]

July 24, 2010

Abstract

We study torture as a mechanism for extracting information from a suspect who may or may not be informed. We show that the optimal use of torture is hindered by two commitment problems. First, the principal would benefit from a commitment to torture a victim 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 are 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 Nageeb Ali, Simon Board, Navin Kartik, Roger Lagunoff, Andy Skrzypacz, and Pierre Yared for useful comments. We also thank seminar audiences at Harvard/M.I.T., U.C. Irvine, Madrid, Michigan, S.E.D. 2010 and U.C.L.A.for comments.

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

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

1 Introduction

A terrorist attack is planned for a major holiday, a few weeks from now. 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 is involved in planning the attack. If we extract his information, 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, as is the question of whether torture should be used to extract his information.[1] Also, the "ticking time bomb" scenario is often invoked in discussions of whether torture is acceptable in extreme circumstances. 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 is "right" to torture because the value of saving many lives outweighs the costs of torture.[2]

This argument for torturing in the first place 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. Second, if after enough resistance we learn that the suspect is likely a low value target, there is an incentive to stop. The

[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. There is also a debate about whether harsh tactics should be used to get information (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..."

suspect knows no useful information. It is better to interrogate another suspect who might be informed. And torture is abhorrent and inflicting it on an uninformed suspect cannot be justified. Both of these commitment problems encourage the informed suspect to resist torture. The first problem means that early revelation leads to yet further revelation under the threat of yet more torture. The second problem means that silence will hasten the cessation of torture. What is the maximal benefit 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. An informed agent has verifiable information - he knows where the bombs are hidden. Also, 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. Each period, the principal decides whether to demand some information from the agent and threaten torture. The suspect either reveals verifiable information or suffers torture. For example, an agent can offer a location for a bomb and the principal can check whether there is in fact a bomb at the reported address. An informed agent can always reveal a true location while an uninformed agent can at best give a false address. This continues until either all of the information is extracted or time runs out. We characterize the unique equilibrium of this game. In equilibrium the informed suspect reveals information gradually, initially resisting and facing torture but eventually he concedes. The value of torture is determined by the equilibrium rate of concession, the amount of information revealed once a concession occurs, and the total length of time that the innocent suspect is tortured along the way.

A number of strategic considerations play a central role in shaping the equilibrium. First, the rate at which the agent can be induced to reveal information is limited by the severity of the threat. If the principal demands too much information in a given period then the agent will prefer to resist and succumb to torture. Second, as soon as the suspect reveals that he is informed by yielding to the principal's demand, he will subsequently be forced to reveal the maximum given the amount of time remaining. This makes it costly for the suspect to concede and makes the alternative of resisting torture more attractive. Thus, in order for the suspect to be willing to concede the principal must also torture a resistant suspect, in particular an uninformed suspect, until the very end. Finally, in order to maintain principal's incentive to continue torturing a resistant suspect the informed suspect must, with positive probability, make his first concession anywhere between the time the principal begins the torture regime to the very end.

These features combine to give a sharp characterization of the value of torture and the way in which it unfolds. Because concessions are gradual and torture cannot stop once it begins, the principal waits until very close to the terminal date before even beginning to torture. Starting much earlier would require torturing an uninformed suspect for many periods in return for only a small increase in the amount of information extracted from the informed. In fact we show that the principal starts to torture only after the game has reached the *ticking time-bomb phase:* the point in time after which the deadline becomes a binding constraint on the amount of information the suspect can be induced to reveal. This limit on the duration of torture also limits the value of torture for the principal.

Because the principal must be willing to torture in every period, the informed suspect's concession probability in any given period is bounded, and this in turn bounds the principal's payoff. In fact we obtain a strict upper bound on the principal's equilibrium payoff by considering an alternative problem in which the suspect's concession probability is maximal subject to this incentive constraint. This bound turns out to be useful for a number of results. For example it allows us to derive an upper bound on the number of periods of torture that is independent of the total amount of information available. We use this result to show that the value of torture shrinks to zero when the period length, i.e. the time interval between torture decisions, shortens. In addition it implies that laws preventing indefinite detention of terrorist suspects entail no compromise in terms of the value of information that could be extracted in the intervening time.

To understand the result on shrinking the period length, note that additional opportunities to torture come at the cost of reducing the principal's temporary commitment power. There are more points in time for the principal to re-evaluate his torture decision and more points where he must be given the incentive to continue. In any time interval, the informed suspect must slow down the rate of information revelation for torture to continue. Over any time interval, we show that as the frequency of decision opportunities increases, the rate of information revelation grinds to a halt. Then, as the frequency of torture opportunities becomes large, the value of torture goes to zero.

This is reminiscent of results like the Coase conjecture for durable goods bargaining but the logic is very different. In our 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 flow cost to the agent limits the amount of information he is willing to reveal in any given segment of real time. As the period length shortens, the principal may torture 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.[3]

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 has a unique equilibrium and hence we obtain sharp bounds of equilibrium payoffs for both players. Unlike the majority of the reputation literature, our model has two long-run players and a terminal date.

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.[4]

We consider two extensions. First to study the use of "enhanced interrogation techniques" we consider 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

[3] A decent, but still not perfect, analogy to bargaining would be the following. 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.

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

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.

Finally, we consider delegating the act of torture to a specialist. Delegation can often solve commitment problems and we have identified two that limit the value of torture. A specialist with a low cost of torture ameliorates one commitment problem: he is willing to continue even if the probability the suspect is informed is quite small. This means the informed suspect can concede information more quickly in equilibrium and the total amount of torture is reduced. On the other hand the specialist cannot commit to limit the torture of the guilty. Indeed, once a suspect starts talking, the ratchet effect applies and the specialist must extract all the information possible in the time remaining. Thus it remains true that once torture begins it must still continue till the terminal date and that the specialist waits until near the end before starting toture. We show that if detaining the suspect is costly and the time horizon is long enough, the value of delegated torture is negative.

Before turning to the formal model, we point out some features that deserve discussion.

Our approach assumes that both players are maximizing their payoffs. There is some evidence that both interrogators and suspects do try to optimize. An Al Qaeda manual describes torture techniques and how to fight them (Post (2005)). American military schools train soldiers how to resist torture. There is also an effort to optimize torture techniques: teachers from military schools helped to train interrogators at the Guantánamo Bay detention center (Mayer (2005)).

We assume that it is costly to inflict torture. 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 (see Alexander and Bruning (2008)). Of course, torture is morally costly. This view begets laws against torture and interrogators may fear prosecution if they use illegal methods. 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. Finally, we assume that information held by the suspect is verifiable. This also gives the best case for torture as a mechanism. If instead all messages were cheap talk, this would reduce the value of torture yet further.

2 Model

There is a torturer (principal) and a suspect (agent). There will be a terrorist attack at time T and the torturer 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 torturer 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 torturer 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.

If the suspect reveals the quantity $y \leq x$ and is tortured for t periods, his payoff is

$$x - y - \Delta t$$

while the torturer's payoff in this case is

$$y - ct.$$

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

2.1 Full Commitment

With full commitment, torture gives rise to a mechanism design problem with hard information which is entirely standard except that there is no individual rationality constraint.

With verifiable information, the only incentive constraint is to dissuade the informed suspect from hiding his information. It goes without saying that a binding incentive-compatibility constraint is a feature of the optimal use of torture.

The torturer demands information $y \leq x$ from the suspect. If he does not reveal this amount of information, he tortures him for $t(y) \leq t$ periods where $t(y) = \frac{y}{\Delta}$. This gives the incentive for the informed suspect to reveal information y at the cost of torturing the uninformed suspect for $t(y)$ periods. The torturer's payoff is

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

and we have the following solution:

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

3 Limited Commitment

We model limited commitment by dividing the real time interval T into periods of discrete time whose length we normalize to 1. There are thus T periods in the game. We assume that the torturer can only commit to torture for a single period. The form of commitment in a given period is also limited. The torturer can demand a (positive) quantity of information and commit to suspend torture in the given period if it is given. Formally, a pure strategy of the torturer 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.

If there are k periods remaining in the game, the maximum cost that can be threatened is $k\Delta$. This is therefore also the maximum amount of information that the informed suspect can be persuaded to reveal. To avoid a trivial case, we assume that $\Delta < x$, i.e. that a single period of torture is not a sufficient threat to induce the agent to divulge all of his information. We measure time in reverse, so "period k" means that there are k periods remaining. But "the first period" or "the last period" means what they usually do.

We begin by defining some quantities. 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 threaten to torture in order to induce revelation of the quantity x (if the principal were able to commit.) Throughout 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 define

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

and define μ_1^* by

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

The function V^1 represents the principal's continuation payoff in period 1 (the last period of the game) when μ is the posterior probability that the (heretofore resistant) suspect is informed. The suspect is threatened with cost Δ and the informed suspect therefore yields Δ. The uninformed suffers torture which costs the principal c.

Next, if μ is a probability that the suspect is informed and q is a probability that he reveals information in a given period, then we define $B(\mu; q)$ to be the posterior probability that the suspect is informed conditional on *not* revealing information in that period. It is given by

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

We define $q_1 = 1$ and a function $q_2(\mu)$ by

$$B(\mu; q_2(\mu)) = \mu_1^* \text{ if } \mu \geq \mu_1^*.$$

i.e.

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

The probability $q_2(\mu)$ will play an important role in the equilibrium. Suppose the suspect has kept silent up to period 2. Then by conceding in period 2 with probability $q_2(\mu)$, he insures that, in the $1-q_2(\mu)$-probability event that he does not concede, the principal will be just willing to continue torturing in the final period.

Now we inductively define functions $V^k(\mu)$ and $q_k(\mu)$ and probabilities μ_k^* as follows.

$$V^k(\mu) = \mu q_k(\mu) \min\{x, k\Delta\} + (1 - \mu q_k(\mu)) \left[V^{k-1}(\mu_{k-1}^*) - c \right]. \tag{2}$$

$$V^k(\mu_k^*) = V^{k-1}(\mu_k^*) \tag{3}$$

$$B(\mu; q_k(\mu)) = \mu_{k-1}^*. \tag{4}$$

These equations will define the value functions and concession probabilities in periods $k = 2, \ldots \bar{k} + 1$ along the equilibrium path. The first task is to show that these quantities are well-defined. Figure 1 illustrates.

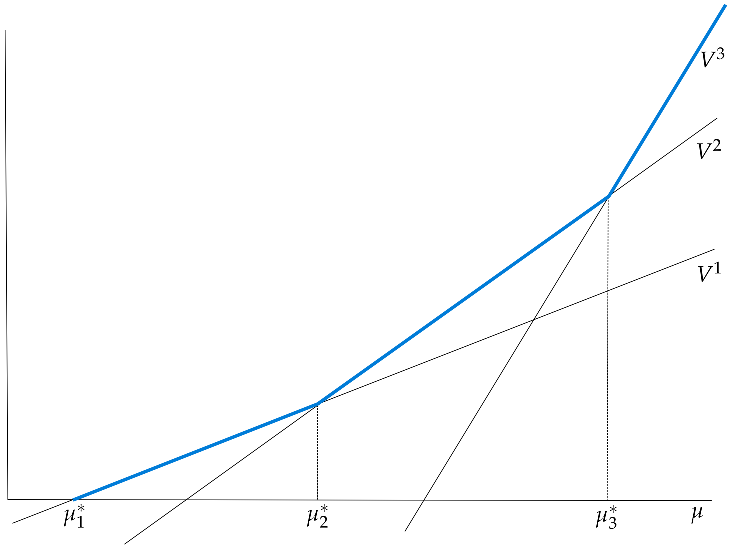

Figure 1: An illustration of the functions V^k and the thresholds μ_k^*. Here $\bar{k} + 1 = 3$. The upper envelope shows the value of torture as a function of the prior μ_0.

Lemma 1. *The above system 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 $k =$ $2, \ldots, \bar{k} + 1$*

We now describe an equilibrium of the game and calculate its payoffs. Subsequently we will show that it is the (essentially) unique equilibrium.

The principal picks the time period $k^* \in \{1, \ldots, \bar{k}+1\}$ that maximizes $V^k(\mu_0)$.[5] The principal delays torture, i.e. sets $y = 0$, until period k^*. In period k^*, with probability 1, the principal demands $y = \Delta$.

In any subsequent period, if the agent has revealed himself to be informed by agreeing to a (non-zero) demand, and if the total quantity x has not yet been revealed, the principal demands Δ (or the maximum amount of information the agent has yet to reveal if that amount is smaller than Δ). If the entire x has already been revealed, the principal stops torturing.

On the other hand, if the agent has resisted torture through period $k < k^*$, then the principal's behavior depends on whether $k = \bar{k}$ or $k < \bar{k}$. (Note that the former case applies only if $k^* = \bar{k}+1$.)

If $k = \bar{k}$ and the agent refused the principal's demand in period $\bar{k}+1$, then the principal randomizes. With probability

$$\rho := \frac{x - \bar{k}\Delta}{\Delta} \tag{5}$$

the principal demands $y = \Delta$, and with the remaining probability the principal does not torture, i.e. sets $y = 0$. On the other hand, if $k < \bar{k}$, and the agent has not yet revealed himself to be informed, the principal, with probability 1, tortures and sets $y = \Delta$.

Next we describe the behavior of the informed agent. (The uninformed agent has no choice to make because he has no verifiable information.) In periods $k = k^*, \ldots, 1$, if he has yet to give in to a positive demand, he will randomize between making his first concession, yielding Δ to the principal, and resisting for another period. The probability of a concession in periods $k < k^*$ is given by $q_k(\mu_k^*)$, and the probability of concession in period k^*, the first period of torture, is $q_{k^*}(\mu_0)$. Finally, in any period in which the informed agent has previously revealed himself to be informed, he agrees, with probability 1, to the principal's demand of Δ.

We have described the following path of play. 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,

[5] Throughout the description we will ignore cases where multiplicity arises due to knife-edge parameter values.

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 $k^* = \bar{k} + 1$.)

In Appendix A, the complete description of equilibrium strategies is given, including off-path beliefs and behavior, as well as the verification of sequential rationality. Here we calculate the payoffs and show the sequential rationality along the path of play.

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 4. 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 4, 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 $V^k(\mu_k^*) = V^{k-1}(\mu_{k-1}^*)$. 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 1, 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 $1, \ldots, \bar{k} - 1$ and to mix in period $\bar{k}$ 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. Finally in period $\bar{k} + 1$, 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 (see Equation 5) this payoff equals x and so the suspect is again indifferent and willing to randomize.

The first main result is that the equilibrium is essentially unique.[6]

Theorem 2. *The unique equilibrium payoff for the principal is*

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

We begin with an observation that plays a key role in the proof and also in subsequent results. Once the suspect reveals some information, say in period k, the continuation game is one of complete information. As shown in the following lemma, in all equilibria of the continuation game

[6] There is some multiplicity in off-equilibrium behavior, and when $k^* = \bar{k} + 1$ it is possible to construct a payoff-equivalent equilibrium in which the torture planned in period $\bar{k} + 1$ alone is moved earlier and behavior at all other periods is the same.

beginning in period $k-1$, the suspect "spills his guts," i.e. he reveals all of his remaining information, up to the maximum torture he can be threatened, $(k-1)\Delta$. The straightforward backward-induction proof is in Appendix B.

Lemma 2. *In any equilibrium, at the beginning of the complete information continuation game with k periods remaining and a quantity $\tilde{x}$ of information yet to be revealed, the suspect's payoff is*

$$-\min\{\tilde{x}, k\Delta\}$$

As we will show in Section 7, this feature represents an additional commitment problem for the torturer. In some instances he would prefer to commit not to extract the maximum amount of information from the suspect. Similar to the "ratchet effect" from the literature on mechanism design without commitment, such a policy cannot be sustained in equilibrium because once the suspect has been revealed to be informed, sequential rationality requires torture to continue.

4 Bounding the Value and Duration of Torture

In this section we develop two important properties of equilibrium which illustrate the limits of torture. First, we establish an upper bound on the principal's equilibrium payoff by considering an additional commitment problem that arises in equilibrium: the principal would like the power to commit to halt torture altogether. In equilibrium this commitment cannot be sustained and so once the torture begins it must continue until the very end. This leads to our second result: the principal will not begin the torture until close to the end. In fact we obtain an upper bound on the number of periods of torture that is independent of the length of the game and the total amount of information available.

Intuitively, if the principal is expected to continue torturing a resistant suspect, the suspect must be conceding at a slow enough rate to ensure that the principal's continuation payoff from torturing is high. On the other hand if the principal had the ability to stop the torture not just for one period, but for the rest of the game, then the suspect could concede with a probability so large as to drive the principal's continuation value to zero. Such an increase in the concession rate would raise the principal's payoff.

In equilibrium however, such a commitment is never credible. Even if the agent were to increase his concession rate and drive the principal's continuation value to zero, the principal could simply pause the torture for a single period. Beginning in the next period the principal's continuation value is positive and he would strictly prefer to resume the torture. This is illustrated in Figure 2 below.

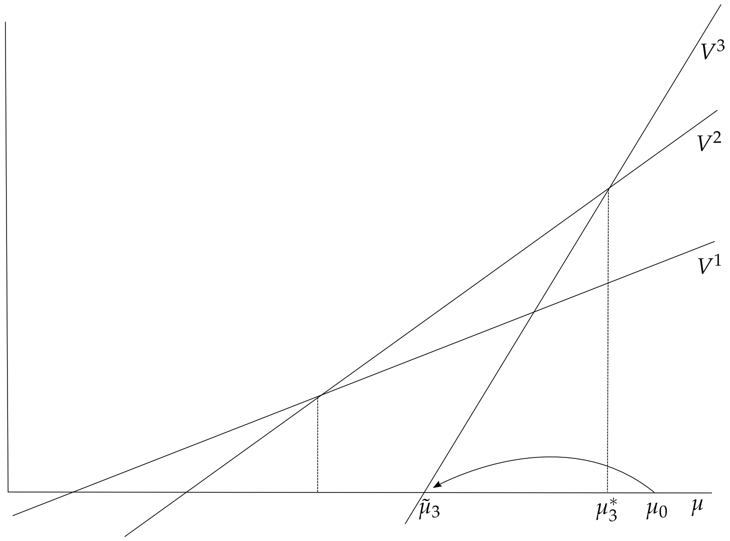

Figure 2: Concession rates would be higher if the principal could commit in period 3 not to torture in periods 2 or 1.

With three periods remaining, at the posterior $\tilde{\mu}_3$ the principal would have a continuation payoff of zero. He would be indifferent between continuing to torture and halting altogether. Being indifferent, he would randomize in such a way as to maintain the suspect's equilibrium payoff. This would enable the suspect to concede with such a probability as to move the principal's posterior from μ_0 to $\tilde{\mu}_3$. In terms of the value of torture, this would improve upon the equilibrium because this represents a higher concession rate than the equilibrium rate which only moves the posterior to μ_3^*. However, without the ability to commit, the principal would prefer to pause torture just in period 3 and then resume in period 2 because his continuation value $V^2(\tilde{\mu}_3)$ is positive.

In addition to illustrating a further commitment problem impeding torture's effectiveness, this observation will provide a useful upper bound on the principal's payoff in equilibrium.

To see this, consider an alternative sequence of functions $\tilde{V}^k(\mu)$ and $\tilde{q}_k(\mu)$ and probabilities $\tilde{\mu}_k$ as follows. First, $\tilde{V}^1(\mu) \equiv V^1(\mu)$, $\tilde{q}_1(\cdot) \equiv q_1(\cdot) \equiv 1$ and $\tilde{\mu}_1 = \mu_1^*$, but for $k \geq 2$,

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

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

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

Following the logic of the equilibrium construction, it is easy to see that these functions define the principal's payoff in an alternative setting in which at each stage the principal either makes a demand $y > 0$ or ends the game. In particular, note that the condition in Equation 7 defines a posterior at which the principal is indifferent between continuing torture and stopping once and for all. As we show in the following theorem, the function $\tilde{V}^k(\cdot)$ gives an upper bound on the principal's equilibrium payoff $V^k(\cdot)$ when there are k periods remaining in the game, and the bound is strict when $k \geq 3$.

Theorem 3. *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 $k \geq 3$.

All proofs in this section are in Appendix C

4.1 Bounding the Duration of Torture

We have shown that once torture begins it must continue until the end. In addition, in order to maintain the principal's incentive to torture, concessions by the suspect must be gradual and spread out over the entire process. Together these properties imply that the longer the principal tortures the slower the concession rate will be. Therefore it is optimal for the principal to wait until very near the end before even beginning to torture. In this section we show how long he will wait.

In particular, we use the results from the previous section to place an upper bound on the number of periods in which there will be torture. Suppose the informed suspect's information x is large and the terminal date T is within the ticking time-bomb phase. The rate at which the agent concedes is then a function of the flow costs of torture c and Δ. These determine the costs and benefits of torture for the principal and hence the rate at which the agent must concede to give the principal the incentive to continue. If the principal begins torture early, the rate of concession is so low that his expected payoff is *negative* given the prior μ_0. The principal instead waits and begins torture well within the terminal date and, for a prior μ_0, there is a bound $K(\mu_0)$ on the duration of torture even if the agent has a large amount of information.

Theorem 4. *Fix the prior μ_0 and define 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*

$$K(\mu_0)\Delta$$

Note that for any given μ_0, the displayed sum converges to infinity in k and therefore $K(\mu_0)$ is finite for any μ_0.

5 Rights Against Indefinite Detention

Theorem 4 implies that, for a fixed torture technology and for a given prior μ_0, there is a time $\bar{T}$ such that *no matter how large x is*, there is never any loss to the torturer to restricting the length of the game to $\bar{T}$. Thus, laws which guarantee prisoner's rights against indefinite detention do not undermine the captor's ability to get the most from torture. Also, Theorem Section 4.1 that there is an upper bound on the amount of information that can be extracted through torture even if the amount of information actually held is arbitrarily large. In particular, the value of torture as a fraction of the first-best value x shrinks to zero as x becomes large[7].

6 Shortening The Period Length

Up to now we have modeled the principal's limited commitment by supposing that decisions to continue torturing are revisited after every discrete torture "episode." The torturer may be able to revisit his strategy almost continuously, reducing his power to commit. To what extent is the value of torture dependent on the implicit power 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.

A given torture technology is parameterized by its flow cost to the suspect (Δ) and to the principal (c.) When the period length is l, this means that the total cost of a single period of torture is $\Delta' = l\Delta$ to the suspect and $c' = lc$ to the principal. In addition, there are now T/l periods in the game and the ticking time-bomb phase consists of $\bar{k}' = x/(l\Delta)$ periods (or the largest integer smaller than that.)

With these modifications in place we can characterize the equilibrium for any $l > 0$ using Theorem 2-Theorem 4. Let $q_k(\mu|l)$ and $V^k(\mu|l)$ and $\tilde{V}^k(\mu|l)$ denote the strategies and value functions obtained for a given l. We are interested in the limit of the principal's payoff as the period length

[7] Since the second-best value (see Theorem 1) is linear in x, the fraction of the second-best value also shrinks to zero.

shortens:

$$\lim_{l \to 0} \max_{k \leq \bar{k}'+1} V^k(\mu_0|l).$$

To obtain a bound, it will be convenient instead to use the upper bound value functions $\tilde{V}(\mu|l)$ as these are homogenous in l. To see this, note for $k = 1, \ldots \bar{k}' + 1$

$$\begin{aligned} \tilde{V}^k(\mu|l) &= \mu q_k(\mu|l) k \Delta' - (1 - \mu q_k(\mu|l)) c' \\ &= l \left[\mu q_k(\mu|l) k \Delta - (1 - \mu q_k(\mu|l)) c \right]. \end{aligned}$$

Then the threshold posterior $\tilde{\mu}_1$ is defined in Equation 7 by

$$\tilde{V}^1(\tilde{\mu}_1|l) = 0$$

so that $\tilde{\mu}_1$ is independent of l. Now by induction, for $k > 1$, $q^k(\mu|l)$ defined in Equation 7 by

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

is independent of l and hence $\tilde{V}^k(\mu|l)$ is linear in l, i.e.

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

for all $k = 1, \ldots, \bar{k}' + 1$.

It follows from Theorem 3 $l\tilde{V}^k(\mu)$ is an upper bound on the principal's continuation payoff when there are k periods remaining and the period length is l. It follows from Theorem 4 that, regardless of the period length, $K(\mu_0)$ is an upper bound on the number of periods of torture and $lK(\mu_0)$ is therefore an upper bound on the real-time duration of effective torture. In particular, the principal's payoff is bounded by $l\Delta K(\mu_0)$. Noting that $K(\mu_0)$ depends only on the the prior μ_0 and the flow costs of torture c and Δ we have established the following.

Theorem 5. *When the time interval between decisions to continue torture approaches zero, the real-time duration of effective torture shrinks to zero and the value of torture shrinks to zero.*

$$\lim_{l \to 0} \max_{k \leq \bar{k}'+1} V^k(\mu_0|l) = 0$$

There are two sources of commitment power for the principal: the endpoint of the game and the discrete intervals of torture. The principal's use of torture leverages both of these. The principal leverages the endpoint by waiting until close to time T before beginning to torture. Nevertheless the results in this section show that the ultimate source of the value of torture is the temporal commitment power given by discrete torture episodes. When these discrete periods are short, the victim's rate of concession slows down to maintain the principal's incentive to torture for more discrete periods. The principal is left with only the terminal date as a source of commitment power and he therefore waits until closer and closer to T before beginning to torture. But this necessarily shrinks his payoff to zero because the threat of torturing for a vanishing length of time can induce revelation of only a vanishing amount of information.

7 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.

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 parameters. This can be seen in a simple example.

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

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

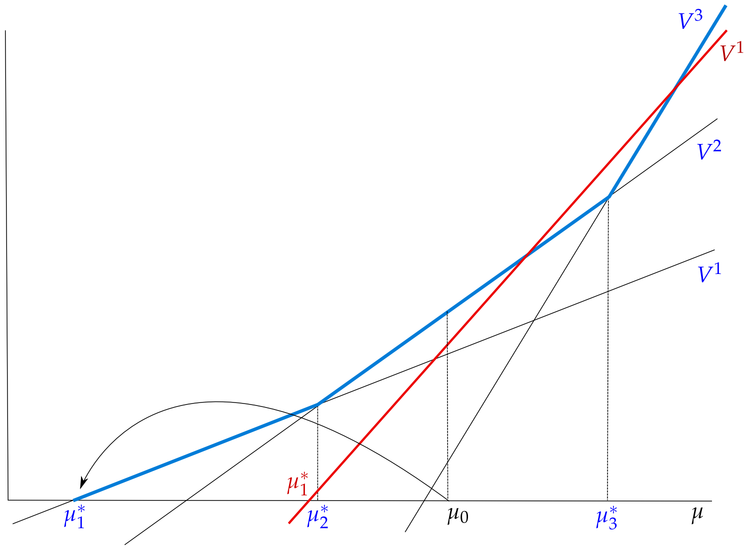

Figure 3: Enhanced interrogation methods undermine the principal's commitment power.

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.

However, because of an important caveat it does not follow that the principal benefits from an array of technologies from which to choose depending on the context. To see why, recall that for any given technology the equilibrium is predicated on the principal's commitment to use that same technology for the duration. Making available the harsher technology 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 3. 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 uninformed 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 3. 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 period 2 and the principal will be forced to wait until the last period and use the harsher technology.

8 Delegation

Torture is limited by commitment problems. One important commitment problem arises because the principal incurs a cost c from torturing. Because of this cost, the principal cannot commit to torture a victim who is almost certain to be uninformed.

This suggests that torture can be more effective when the task of carrying out the torture is delegated to a specialist. In the model, the period-by-period decision whether to continue torture is governed by the torturer's perceived cost of torturing. If the torturer 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 victim. 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 with a cost c' arbitrarily close to zero[8], could effectively commit to torture innocent suspects and thereby extract immediately the entire quantity x of information from the informed. In this sense, delegation to a specialist can alleviate one of the commitment problems inherent in torture.

But other problems are not as easily solved. We have shown, for example, that regardless of the value of c, torture does not commence until the ticking time-bomb phase, a time interval x/Δ that is independent of c. Thus, even a specialist with a low c will delay torture, possibly for a long time, and this itself could be costly.

To illustrate this, let's assume that there is a flow cost $r > 0$ incurred every period that the victim is detained. These detention costs capture the cost of housing and feeding the victim, the political costs of detaining him without trial, and the opportunity cost of not freeing up space for the next victim. We can analyze the model in which the specialist has a perceived cost c' of torturing and a perceived cost r' from detaining the victim. The principal incurs the true flow costs c from torture and r from detention. The principal decides at the beginning of the game whether to employ the specialist. If the specialist is employed then the victim is detained and the game is played using the specialist's c'. At any point in time the specialist can end the game, stopping torture and avoiding further detention costs. The principal's payoff is evaluated based on the total information extracted, the total amount of torture (evaluated using the true cost c) and the total detention costs. If the specialist is not employed, then the game ends immediately and the principal's payoff is zero.

The principal would like the specialist to commence, and conclude, the torture earlier rather than later in order to save detention costs. However, this cannot happen in equilibrium. For example, suppose that the specialist planned to carry out his torture in the first x/Δ periods rather than the last. Consider what happens in the last period of planned torture. There must be a positive probability that the informed will make his first concession in that last period, otherwise the torture would have stopped earlier.

[8] Choosing a specialist with $c = 0$ is problematic because such a specialist would have no incentive to ever stop torturing. This would create multiple equilibria including equilibria in which there is too much torture. Choosing a specialist with a taste for torture $c < 0$ is even worse.

That first concession yields only Δ out of the total quantity of information x. Since there are many periods remaining before the exogenous terminal date, the "spill your guts" logic applies and the specialist will continue to torture in order to extract additional information. But this undermines the informed victim's incentive to begin conceding. If instead he remained silent, the torture would have ended as planned and his payoff would be only $-\Delta$.

The equilibrium path of torture maintains incentives only because the last period of torture is the true, exogenous, end of the game. That terminal point solves the second key commitment problem for the principal: the promise to limit the torture of an informed victim. Moreover, the problem cannot be solved by judiciously choosing the specialist's perceived cost of detention, r'. If r' is smaller than Δ then the specialist will choose to incur the detention costs and continue torturing. If r' is greater than or equal to Δ then the specialist would have stopped the torture one period earlier, unraveling the previous period's incentives. Indeed, regardless of the values of c' and r', this second commitment problem remains: once torture begins, it must continue until the exogenous terminal date. Thus an upper bound for the value of torture, when delegated to a specialist, is

$$\mu x - rT$$

which can be negative when the ticking time-bomb is far enough in the future.

9 Conclusion

Under the threat of an imminent attack, a simple cost-benefit calculation recommends torture: the cost of torture pales in comparison to the value of lives saved by using extracted information. We show that this logic depends crucially on the assumption that it is possible to commit to a torture incentive scheme. When the principal can revisit his torture strategy at discrete points in time, the informed agent must concede 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.

Torture can be contrasted with alternative mechanisms. One possibility is to pay suspects for information. At first glance this mechanism appears strategically equivalent to torture, where paying a dollar is equivalent to reducing torture by one unit. Note however that a "carrot" mechanism using money avoids one of the commitment problems inherent in torture. It is easy to credibly commit not to pay the uninformed. If torture is also an available instrument, a carrot mechanism encounters the same difficulty as a mild torture technology when an enhanced interrogation technique is available. Once the suspect starts talking for payment of a reward, the principal can switch and threaten him with torture unless he gives up information for free. This causes the carrot mechanism to unravel and the same issues that we study come up again.

Finally, we have made some simplifying assumptions to keep our model tractable. For example, 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. This scenario creates some intriguing possibilities when there is limited commitment. Perhaps a middle level target starts talking immediately in equilibrium while a high level target concedes slowly and pretends to be uninformed. This issue and many others await further 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.

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

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.

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 (1985): "Foundation of dynamic monopoly and the coase conjecture," .

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.

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.

Mayer, J. (2005): "The Experiment: The military trains peopole to withstand interrogation. Are those methods being misused at Guantánamo?," *The New Yorker*, p. 60.

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.

Sobel, J., and I. Takahashi (1983): "A multistage model of bargaining," *The Review of Economic Studies*, 50(3), 411–426.

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

A Full Description And Verification of the Equilibrium

Proof of Lemma 1. By Equation 1 and Equation 4,

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

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

We have already described the behavior on-path. Now we describe the behavior after a deviation from the path. If the victim 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 torturer ignores any deviations by the victim along histories where the victim has already revealed information.

If no information has been revealed yet, then behavior after a deviation by the torturer depends on whether $k^* < \bar{k} + 1$ or $k^* = \bar{k} + 1$ and on the value of the current posterior probability μ that the victim 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 victim refuses any demand y greater than Δ. On the other hand if the torturer deviates and asks for $0 < y \leq \Delta$, then the victim 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 victim 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 victim 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 torturer 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 victim refuses with probability 1. If $y \leq x$ then t then the deviation alters the continuation strategies in two ways. First, the informed victim 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 (see Equation 5.) Effectively, this deviation has just shifted the torture that would have occurred in period $\bar{k} + 1$ to the earlier period k.

B Proof of Theorem 2

Proof of Lemma 2. First suppose that $k = 1$ so that there is a single period remaining and assume that the victim 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 torturer is threatening to inflict Δ on the victim. If $y > \Delta$ the victim will refuse, if $y < \Delta$, the victim strictly prefers to reveal y and if $y = \Delta$ he is indifferent. The unique equilibrium is for the torturer to ask for $y = \min\{\tilde{x}, \Delta\}$ and for the victim to reveal y. This gives the victim 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 victim and $\min\{\tilde{x}, (k-1)\Delta\}$ for the torturer and assume that there are k periods remaining and $\tilde{x}$ has yet to be revealed. Suppose the victim is asked in period k to reveal $y \leq \min\{\tilde{x}, \Delta\}$ or else endure torture.

If the victim 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 victim 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 torturer asks for $\min\{\tilde{x}, \Delta\} - \varepsilon$, sequential rationality requires that the victim complies. By the induction hypothesis this leads to a total payoff of $\min\{\tilde{x}, k\Delta\} - \varepsilon$ for the torturer. Since $\min\{\tilde{x}, k\Delta\}$ is the maximum payoff for the torturer consistent with feasibility and individual rationality for the victim, it follows that all equilibria must yield $\min\{\tilde{x}, k\Delta\}$ for the torturer.[9] Any strategy profile which gives this payoff to the torturer must involve maximal revelation ($\min\{\tilde{x}, k\Delta\}$) and no torture. Thus, all equilibria give payoff $-\min\{\tilde{x}, k\Delta\}$ to the victim. □

The following simple implication of Bayes' rule will be useful.

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

$$q + (1-q)q_k(B(\mu; q)) = q_k(\mu). \tag{9}$$

Proof. The equality follows immediately from the fact that $B(\mu; \cdot)$ applied to either side yields μ^*_{k-1}. Intuitively, no matter what the probability of revelation in period $k+1$, the function q_k adjusts the probability of revelation in period k so that the posterior probability of an informed victim conditional on no revelation in either period will equal μ^*_{k-1}. On the left-hand side the probability of revelation in period $k+1$ is q and on the right-hand side it is zero.

An explicit calculation follows.

[9] 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 $\tilde{x}$.

$B(\mu;\cdot)$ applied to the right-hand side of (9) gives μ^*_{k-1}. Applying $B(\mu;\cdot)$ to the left-hand side gives

$$\begin{aligned} B(\mu; q + (1-q)q_k(B(\mu;q))) &= \frac{\mu\,(1 - [q + (1-q)q_k(B(\mu;q))])}{1 - \mu\,[q + (1-q)q_k(B(\mu;q))]} \\ &= \frac{\frac{\mu(1-q)}{1-\mu q}\,[1 - q_k(B(\mu;q))]}{1 - \frac{\mu(1-q)q_k(B(\mu;q))}{1-\mu q}} \\ &= \frac{B(\mu;q)\,[1 - q_k(B(\mu;q))]}{1 - B(\mu;q)q_k(B(\mu;q))} \\ &= B(B(\mu;q); q_k(B(\mu;q))) \\ &= \mu^*_{k-1}. \end{aligned}$$

The Lemma follows from the fact that $B(\mu;q)$ is invertible. □

Proof of Theorem 2. Because Lemma 2 characterizes continuation equilibria following a concession, the analysis focuses on continuation equilibria following histories in which the victim has yet to concede, and the posterior probability of an informed victim 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 $y > 0$.

The proof has three main parts. We first consider continuation equilibria starting in a period $k \leq \bar{k}$ in which there is torture in period k. We show that the unique continuation equilibrium payoff for the principal is $V^k(\mu)$. The second step is to consider continuation equilibria starting in a period $k > \bar{k}$. We show that if there is torture in period k then k is the only period earlier than $\bar{k}$ in which there is torture and the principal's payoff is $V^{k+1}(\mu)$. The final step uses these results to show that in the unique equilibrium of the game, the principal begins torturing in the period k which maximizes $V^k(\mu_0)$.

For the first step, we will 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 victim 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 positibe 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 Lemma 2 that $y \leq \Delta$. For if the informed victim 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 victim 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 $B(\mu; q) < \mu_{k-1}^*$ and the induction hypothesis, there will be no torture in period $k-1$ if the victim resists in period k. This means that a resistant victim has a payoff no less than $-(k-1)\Delta$. But if the victim concedes in period k, by Lemma 2, his payoff will be $-y-(k-1)\Delta$. The informed victim 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. $B(\mu; q) > \mu_{k-1}^*$ then by the induction hypothesis the continuation equilibrium after the victim resists gives the victim a payoff of $-(k-1)\Delta$ for a total of $-k\Delta$. But conceding gives $-y-(k-1)\Delta$ by Lemma 2 and thus the victim strictly prefers to concede, a contradiction since $q < q_k(\mu)$ requires that the victim weakly prefers to resist.

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

$$k\Delta\mu q + (1-\mu q)\left[V^{k-1}(B(\mu;q)) - c\right] \tag{10}$$

(applying Lemma 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 victim must concede with probability $q_k(\mu)$.

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

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

Substituting into Equation 10, 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.

$$\begin{aligned}\mu q\left[k\Delta + V^{k-2}(\mu^*_{k-2}) + 2c\right]&\\ + (1-\mu q)Z(q)\left[(k-1)\Delta - V^{k-2}(\mu^*_{k-2}) + c\right]&\\ + V^{k-2}(\mu^*_{k-2}) - 2c&\end{aligned} \quad (11)$$

Now, by Lemma 3,

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

If we multiply both sides by μ

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

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

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

we obtain

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

or

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

Thus, the coefficients in Equation 11, μq and $(1-\mu q)Z(q)$ sum to a constant, independent of q. It follows that the principal's payoff is 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 victim 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 victim'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.

For the second step, begin by considering continuation equilibria beginningin period $\bar{k}+1$. Then we can follow the same argument from the preceding inductive step to show that the principal demands $y = x - \bar{k}\Delta$, the informed agent concedes with probability $q_{\bar{k}+1}(\mu)$ and then subsequently (by Lemma 2) yields the entire quantity x. Furthermore:

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

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

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

We now consider by induction on j continuation equilibria beginning in period $\bar{k}+j$. In this case we show that the conclusions of three claims above are unchanged:

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

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

3. If $\mu < \mu^*_{\bar{k}+1}$ then there is no equilibrium with a positive probability of torture in period $\bar{k} + j$.

(In other words, equilibria with torture in period $\bar{k} + j$ are payoff equivalent to equilibria with torture in period $\bar{k} + 1$.)

Suppose the claim is true for $j \geq 1$. Consider an equilibrium in which torture begins in period $\bar{k} + j + 1$. If there is no other period of torture between $\bar{k} + j + 1$ and $\bar{k}$, then the equilibrium is payoff equivalent to one in which the torture begins instead in period $\bar{k} + 1$ and we are done.

We will now show that there can be no other period of torture between $\bar{k} + j + 1$ and $\bar{k}$. Let z be the earliest such period in which there is torture. If the informed victim concedes with positive probability in period $\bar{k} + j + 1$ then his total payoff from conceding is $-x$ by Lemma 2. On the other hand, his total payoff from resisting is $-\Delta - \tau$ where τ is some element of $[\bar{k}\Delta, x]$. This follows from the induction hypothesis since $[\bar{k}\Delta, x]$ is the set of possible continuation values for the victim if he has yet to concede by period z. We can rule out $\tau = x$ because then the victim would strictly prefer to concede. That is impossible because then the posterior after resistance in period $k + j + 1$ would be 0 and there would be no torture in period z. So $\tau \in [\bar{k}\Delta, x)$ which implies by the induction hypothesis that the posterior in period z must be $\mu^*_{\bar{k}+1}$. Therefore the informed victim concedes in period $j + k + 1$ with the probability q such that $B(\mu; q) = \mu^*_{\bar{k}+1}$, call it $q_{\bar{k}+2}(\mu)$. Note that $q_{\bar{k}+2}(\mu) < q_{\bar{k}+1}$.

The principal's payoff is

$$\mu q_{\bar{k}+2(\mu)} x + (1 - \mu q_{\bar{k}+2}(\mu)) \left[V^{\bar{k}+1}(\mu^*_{\bar{k}+1}) - c \right].$$

Since $V^{\bar{k}+1}(\mu^*_{\bar{k}+1}) = V^{\bar{k}}(\mu^*_{\bar{k}+1})$, this is strictly smaller than $V^{\bar{k}+1}(\mu)$. This is impossible in equilibrium because then the principal would prefer not to torture in period $\bar{k} + j + 1$ and instead begin the torture in period $\bar{k} + 1$ and obtain his continuation equilibrium payoff of $V^{\bar{k}+1}(\mu)$.

That concludes the second step 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 $V^j(\mu_0)$ and 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 4

Proof of Theorem 3. 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 is bounded by

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

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

Given the definition of $\tilde{V}^{k-1}(\cdot)$ and writing $Z(q) = B(\mu; q)\tilde{q}_{k-1}(B(\mu; q))$ we can write the maximand as

$$\mu q \min\{x, k\Delta\} + (1 - \mu q) \left[Z(q) \min\{x, (k-1)\Delta\} - c\,(1 - Z(q)) - c \right]$$

which can be re-arranged as follows.

$$\mu q \left[\min\{x, k\Delta\} + 2c \right] + (1 - \mu q) Z(q) \left[\min\{x, (k-1)\Delta\} + c \right] - 2c \qquad (12)$$

By Lemma 3 (and the same manipulations as in the proof of Theorem 2) the maximand is strictly increasing in q and therefore since $q_k(\mu) < \tilde{q}_k(\mu)$ we have

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

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 $\tilde{V}^k(\mu)$. □

Proof. If the principal begins torturing in period k, then his payoff $V^k(\mu_0)$ must be non-negative. By Theorem 3 $\tilde{V}^k(\mu_0) \geq V^k(\mu_0) \geq 0$ and therefore $\mu_0 \geq \tilde{\mu}_k$. 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 $\tilde{V}^j(\tilde{\mu}_j)$,

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

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

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

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 Theorem 3. 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. □