Torture in Counterterrorism: Agency Incentives and Slippery Slopes

Hugo M. Mialon, Sue H. Mialon, and Maxwell B. Stinchcombe[1]

January 10, 2010

Abstract

We develop a model of counterterrorism to analyze the effects of allowing a government agency to torture terrorist suspects. We find that legalizing torture in high evidence cases has offsetting effects on agency incentives to counter terrorism by means other than torture. It increases these incentives because other efforts may increase the probability of having high enough evidence to warrant the use of torture if other efforts fail. However, it also lowers these incentives because the agency might come to rely on torture to avert attacks. If the latter effect dominates, legalizing torture in high evidence cases can reduce security and increase the probability of terrorist attack. Moreover, it can increase agency incentives to torture even in low evidence cases, leading to a "slippery slope." (JEL K4, D8, H1)

Keywords: Terrorism, Torture, Law, Agency, Deskilling, Commitment, Complementarity, Enforcement, Security, Slippery Slope

[1] Hugo Mialon and Sue Mialon, Department of Economics, Emory University, Atlanta GA 30322-2240 (hmialon@emory.edu and smialon@emory.edu). Maxwell Stinchcombe, Department of Economics, University of Texas at Austin, Austin TX 78712-1173 (maxwell@eco.utexas.edu). We are extremely grateful to Stephen Clark, Martin Dumav, Amy Farmer, Andrew Francis, Louis Kaplow, Tilman Klumpp, Preston McAfee, Tim Mathews, Erik Nesson, Joshua Robinson, Paul Rubin, Louis Seidman, Steve Shavell, Jeroen Swinkels, Kathy Zeiler, and seminar participants at Emory University, Harvard Law School, and the 2009 Southern Economic Association Meeting for wonderfully helpful comments.

1. Introduction

Since the September 11, 2001 terrorist attacks on the U.S., a number of legal authorities, political authorities, and poll results have favored the use of torture in counterterrorism. Many arguments in favor of the use of torture begin with, or prominently feature, some variant of a "ticking bomb" scenario in which torturing one suspect leads, with near certainty, to saving many lives. However, if a ticking bomb scenario arises, it is because other preventive efforts have failed. In this paper, we analyze the effects of legalizing torture that arise through the "portfolio" effects linking torture and other kinds of counterterrorism efforts.

1.1 The Push For Torture

In January of 2002, as Assistant Attorney General in the U.S. Department of Justice, John Yoo recommended that the White House withdraw its recognition of the rules prohibiting torture imposed by the Geneva Conventions.[2] In August of 2002, as head of the Office of Legal Counsel at the Justice Department, Jay Bybee recommended that the prohibition on torture be reserved only for the infliction of extreme pain associated with death or organ failure.[3] In an influential book, Dershowitz (2002) proposed a system of judicial warrants for the use of non-lethal torture for counterterrorism purposes in certain limited circumstances.

In 2004, Senator Charles Schumer publicly rejected the idea that torture should never be used.[4] In 2006, Congress passed, and the President signed into law, the Military Commissions Act (Public Law No. 109-366), which limits judicial review for detainees to challenge

[2] Memorandum from John Yoo, Deputy Assistant Attorney General, and Robert J. Delahunty, Special Counsel, to William J. Haynes II, General Counsel, Department of Defense (Jan. 9, 2002).
[3] Memorandum from Jay S. Bybee, Assistant Attorney General, Office of Legal Counsel, U.S. Department of Justice, to Alberto R. Gonzales, Counsel to the President, Standards of Conduct for Interrogation Under 18 U.S.C. §§2340-2340A (Aug. 1, 2002).
[4] Federal Government's Counterterrorism Efforts: Hearing Before the Sen. Judiciary Subcommittee, 108th Cong. (2004) (statement of Sen. Charles Schumer, Member, S. Judiciary Committee).

their treatment. The Act also permits admission of evidence obtained through torture under certain circumstances, at least if the interrogation in question occurred prior to December 2005 (see Martinez, 2007). In 2007, during his confirmation hearings, Attorney General Michael Mukasey refused to state that waterboarding (a recognized form of torture that subjects suspects to the terror of drowning) is illegal.[5] In 1987, the Supreme Court of Israel consented to the use of torture to stop attacks by Palestinian terrorists (Imseis, 2001, and Strauss, 2003). In a 1999 decision, the Court returned to a ban on torture, but the ban is not absolute and has not consistently been enforced as discussed by Imseis (2001).

The Economist (2006) reported results of public opinion polls about torture. In a BBC poll of 27,000 people in 25 countries, 33 percent of people polled, 36 percent of Americans polled, and 46 percent of Israelis polled said that "some degree of torture is permissible." In a recent poll by the Pew Research Foundation (2009) of 1,303 Americans, 44 percent of people polled thought that "torturing terrorist suspects is often or sometimes justified."

1.2 The Ticking Bomb Scenario

Most arguments in favor of legalizing torture are presented as variants of the following mass terrorism scenario.[6] Suppose the government learns that terrorists are planning an attack in a populated area. If the attack succeeds, many people could die. The government has in custody a suspect who may know about the attack but refuses to cooperate. The government can perhaps force the suspect to reveal what he knows through torture. The suspect could survive the torture, and the information extracted could save many lives. The government does not, at that time, have other means of stopping the attack.

In at least some versions of this scenario, cost-benefit analysis strongly suggests that

[5] Executive Nomination: Hearing on the Nomination of Michael B. Mukasey To Be Attorney General of the United States Before the S. Comm. on the Judiciary, 110th Cong. (2007). See Egan (2007 a,b)
[6] See e.g. Dershowitz (2002), Strauss (2003), Luban (2005), and Bagaric and Clarke (2006).

allowing the government to use torture on the suspect is the socially efficient policy. In this paper, we examine this efficiency argument more carefully. There are unintended consequences to allowing torture, and those consequences vary with the degree of evidence required to make torture legal.

More specifically, if a ticking bomb scenario arises, it reflects a failure of other preventive efforts.[7] Assuming that it is possible for situations to arise in which the cost-benefit calculations described above favor torture, we develop a model of counterterrorism to evaluate the overall effects of legalizing torture. Throughout, we maintain the assumption that there is a fundamental agency problem: the agency tasked with counterterrorism does not fully internalize the social damages from a terrorist attack and cares less about protecting the rights of individuals than does society.

1.3 When The Agency Obeys Directives

For the first part of our analysis, we assume that the agency obeys directives on torture policy. We show that allowing the agency to torture when evidence of guilt is high has two opposing effects on the agency's incentives to counter terrorism by means other than torture. First, it tends to reduce these incentives because it ameliorates a situation in which other efforts have failed–a decommitment effect. Second, it tends to increase these incentives because other efforts improve the chances of gaining enough evidence to warrant the use of torture if the other efforts fail–a complementarity effect. We also show that allowing torture in a broader range of cases lowers the complementarity effect.

When the decommitment effect dominates, legalizing torture reduces the agency's other

[7] For domestically based attacks, such efforts include tracking of materials used in bomb-making, restrictions on bomb-making activities, increased security at likely targets, and baggage and cargo screening at airports. For extraterritorial sources, such efforts include hiring, training and paying attention to analysts fluent in language and culture, cultivating allies, and bilateral or multilateral cooperative international policing.

preventive efforts, and can thereby reduce security as well as increase the probability of torturing the innocent. In this case, we have a formalization of the observation in Rejali (2007) that reliance on torture typically makes an agency sloppier in its other preventive work and leads to agency "deskilling." In the longer run, it might reduce investment on the development of alternative technologies and prevention techniques.[8]

1.4 The Enforcement Problem

The core of the Dershowitz (2002) argument is that agencies do not obey directives, that torture happens even though it is illegal, and that an enforced system of judicial warrants could bring this under control, resulting in less risk of torturing the innocent. In the second part of our analysis, we extend our model to encompass the possibility that the agency is willing to disobey directives on torture at the risk of legal sanction.[9] In our extended model, the agency can choose whether to use torture even when torture is not allowed, and it faces potential punishment if it uses torture illegally. If torture is legal in high evidence cases, the

[8] Rejali (*op. cit.* Chap 21 and 22) documents the deskilling effect across a large number of 19'th and 20'th century instances. For example, there is evidence that the Gestapo's suppression of the Resistance in World War II was far more effective when it relied on informers and careful interrogation before it turned extensively and "unprofessionally" to torture.

In the French-Algerian war, French army units that tortured became insubordinate to central army authority and abandoned basic police techniques. In one instance, going directly to torture rather than checking the personal effects of an apprehended suspect allowed an Algerian resistance bomb factory to be safely relocated. The radical part of the Algerian resistance movement followed a policy of identifying members of the moderate opposition when tortured, and because the French army did not check the veracity of what was revealed under torture, it wiped out the moderate opposition.

For a more recent example, when Abdul Hakim Murad was arrested by Filipino police in 1995 with fake passports, bomb-making materials, and an encrypted computer, police tortured him instead of trying to decrypt the computer. He revealed little specific information under torture, but when the CIA decrypted his computer years later, it revealed detailed information about Al Qaeda plots to blow up planes in the US, down to specific procedures and flight schedules.

The CIA's unedited *Human Resources Exploitation Training* manual summarizes the deskilling effect of torture with "The routine use of torture ... corrupts those that rely on it as the quick and easy way out" (See the National Security Archives at website http://www2.gwu.edu/~nsarchiv/NSAEBB/NSAEBB122/).

[9] In wartime and on foreign grounds, the risk that torture will be used illegally may be high, as evidenced by the documented reports of sadistic torture at the Abu Ghraib and Guantanamo Bay detention facilities of the U.S. Military (Fay, 2004).

agency only faces a potential punishment if it uses torture in cases where evidence is low.

In this extended context, we find that legalizing torture in high evidence cases through a torture warrant system has three effects on the agency's non-torture efforts. First, it tends to reduce such efforts because it eliminates the agency's cost of using torture in high evidence cases–the decommitment effect. Second, it tends to increase such efforts because the agency thereby increases the chances of obtaining high evidence if the efforts fail to stop an attack, in which case the agency can torture with impunity–the complementarity effect. Third, it tends to reduce such efforts because the agency thereby increases the chances that it has high evidence even if there turns out to be no attack, in which case the agency can escape punishment for using torture on innocent individuals–the decomplementarity effect.

If the complementarity effect dominates the decommitment and decomplementarity effects, then legalizing torture in high evidence cases increases the agency's other efforts, which increases the accuracy of the agency's evidence, thereby potentially reducing the probability that an innocent person is tortured. This case supports the Dershowitz argument that an open warrant system might actually increase incentives to obey the law and reduce torture of the innocent. However, if the decommitment and decomplementarity effects dominate the complementarity effect, then legalizing torture in high evidence cases reduces the agency's other efforts and thereby increases the probability that an innocent person is tortured.

1.5 Slippery Slopes

We also find that legalizing torture in high evidence cases can lead to an increase in its use in other cases, i.e., a "slippery slope." Intuitively, this involves the endogeneity of the quality of information. If the decommitment effect dominates, legalizing torture in high evidence cases reduces the agency's efforts to counter terrorism by means other than torture, which in turn reduces the quality of the information on which the agency bases its decision to use

torture if the other efforts fail, increasing agency incentives to use torture in other cases.

This mechanism differs from the three basic variants of the slippery slope arguments that we have found in the literature: utility change, cost change, and somewhat related bureaucratic structure arguments. Volokh (2003, p. 1077) elegantly summarizes the utility change arguments as "the normative power of the actual." In more pedestrian language, a society that allows torture will perhaps come to see nothing wrong with it.

Volokh (2003) and Rizzo and Whitman (2004) provide a detailed examination of cost-based slippery slope mechanisms in legal policymaking. These involve one decision lowering the cost or otherwise increasing the incentives to make another linked decision.[10] In our context, if society pays the cost of training and supporting professional torturers, then the lower marginal cost of torture can lead to more frequent torture.

Posner (2002, p. 30) argues that if "... rules are promulgated permitting torture in defined circumstances, some officials are bound to want to explore the outer bounds of the rules. Having been regularized, the practice will become regular." Sobel (2000, 2001) provides insightful models of declining standards that may bear on the worry that any chosen evidence standard for torture may be prone to slip over time, perhaps by the accretion of precedents set by judges who are more sympathetic to agency arguments for torture.

Nonetheless, it is possible that a legal standard for torture would not slip if torture were legalized, just as the legal standard for capital punishment does not seem to have slipped after capital punishment was legalized (Bagaric and Clarke, 2006). However, according to the slippery slope mechanism that we identify here, even if the legal standard for torture were not to slip, legalizing torture in certain circumstances could still entail a slippery slope in that it could increase illegal torture in other circumstances.

[10]For example, street corner cameras to prevent crime make the cost of government tracking of all citizens using face recognition software much lower.

1.6 Endogenous Terrorism

We derive the above results assuming that terrorist activity is exogenous. In the third part of our analysis, we endogenize terrorist activity and examine how legalizing torture in high evidence cases affects the probability of a terrorist attack. We consider a model that is identical to the basic model except that individuals choose whether or not to initiate a terrorist action. We find that, if the decommitment effect dominates the complementarity effect, then legalizing torture in high evidence cases indirectly increases the probability of attack by reducing the agency's preventive effort and thereby increasing the probability that an attack would succeed. We also find that legalizing torture in high evidence cases directly increases the probability of attack by increasing the expected payoff of attacking relative to the expected payoff of not attacking for any given level of preventive effort by the agency if torture is not too highly effective and the costs of being tortured are sufficiently lower for innocent individuals than for individuals who become terrorists.

Lastly, we argue that legalizing torture may also increase the probability of attack because it may serve as a signal that a government is illegitimate and thereby increase the expected benefits to individuals of committing terrorist acts against the government.

1.7 Organization

Section 2 discusses models of torture in the literature. Section 3 presents our basic model of counterterrorism when the agency obeys directives on torture policy. Section 4 extends the model to consider the enforcement problems that arise from the agency being willing to run the risk of legal sanction. Section 5 endogenizes terrorist activity. Section 6 summarizes and discusses avenues for future work.

2. Models of Torture in the Literature

There is a growing literature on the economics of terrorism (see Enders and Sandler, 2004 and 2005, Sandler and Siqueira, 2006, Siqueira and Sandler, 2007, Garoupa, Klick, and Parisi, 2006, and Berman and Laitin, 2008). However, this literature has not considered the use of torture in counterterrorism. If one considers avoidance of torture an individual right, the small but growing literature on the economics of individual rights is relevant.[11] Seidmann and Stein (2000), Mialon (2005), Leshem (2009), and Wickelgren (2010) examine the right to silence. These papers analyze the effects of preventing adverse inferences from a suspect's silence during interrogation, but they do not analyze the arguably more fundamental right against torture in interrogation.

We have found very few formal analyses of torture in the literature. Wantchekon and Healy (1999) analyze torture as a game of incomplete information between a state, a torturer, and a victim. The state chooses whether to sanction torture to extract information. If it sanctions torture, it hires an agent to carry out the torture. Then the victim and torturer alternately choose, respectively, how much to reveal and how much to torture. The victim and torturer do not know each other's type. The victim is either guilty or innocent and either weak or strong. The strong victim is difficult to intimidate. The torturer can be professional or sadistic. A professional incurs a cost from torturing and therefore does not torture unless he thinks that this will yield information. A sadist derives a personal benefit from torture and therefore tortures even if he thinks that it will not extract information. The authors demonstrate that if the state sanctions torture, then torture will be carried out with positive probability in equilibrium because even a professional torturer might torture

[11] Mialon and Rubin (2008) provide a summary and synthesis. Atkins and Rubin (2003), Garoupa (2005), and Mialon and Mialon (2008) analyze issues related to the right against unreasonable searches. Shavell (1991), Andreoni (1991), and Persson and Siven (2007) analyze issues related to the right against cruel and unusual punishment.

to test whether the victim is weak or strong and even a weak victim might hold out to test whether the torturer is professional or sadistic.

Chen, Tsai, and Leung (2009) analyze a model of judicial torture. In their model, a defendant, who is either innocent or guilty, chooses whether to confess. If he confesses, he is convicted. If he does not confess, the judge conducts an independent investigation that produces evidence about whether the defendant is guilty and, if torture is allowed, chooses whether to torture the defendant to force confession and conviction. If the judge does not torture the defendant, the defendant is released. The authors compare outcomes under a system where the judge is allowed to torture and under a system where the judge cannot torture and thus chooses whether to convict based solely on the evidence about whether the defendant is guilty. The authors demonstrate that the evidence-based system dominates the torture system if the independently-produced evidence is sufficiently accurate. They then employ this result to explain the historical decline in the use of torture in judicial proceedings with the historical advancement in independent investigation technologies.

Chen, Chou, and Tsai (2009) model judicial torture as a war of attrition. They show that judicial torture may occur because of the magistrate's uncertainty about the suspect's limit of pain endurance and the suspect's uncertainty about the magistrate's limit on pain infliction. Moreover, they show how the magistrate's decision to use torture depends on a number of factors, including the magistrate's concern for type II errors.

In our paper, we do not model judicial torture but rather model torture in counterterrorism, and we do not focus on the strategic interaction between torturer and suspect but instead focus on the effects of legalizing and regulating torture on the behavior of the counterterrorism agency. Unlike the above papers, we consider the implications for security and torture of the innocent of agency problems and problems enforcing directives on torture.

Stephenson (2007) provides a general analysis of the effects of bureaucratic costs on agency expertise. He finds that increasing the costs to an agency of taking a regulatory action decreases the agency's incentive to invest in learning more about the benefit of the proposed action if the agency would not have taken the action if it did not learn more about the action's benefit. The agency's incentive to learn more about the action's benefit, however, increases if the agency would have taken the action even if it did not learn more about the benefit. Our model has a few similarities with Stephenson's model. In our model, the agency faces costs from using torture illegally, and these costs affect the agency's effort to counter terrorism by means other than torture. However, the context we analyze is different and we focus on a different set of issues, including the possibility of slippery slopes.

3. When the Agency Obeys Directives

We begin with a description of the basic model of a government agency (e.g., the CIA, the FBI, or DOD military trainers) and an individual who may have initiated a terrorist action. Under the assumption that the agency obeys directives about torture, we compare outcomes, in terms of agency behavior and social welfare, under three scenarios: torture is illegal; torture is illegal except in the face of strong evidence of suspect guilt; and torture is legal. The next section extends the model to study the enforcement problem that arises if the agency may choose, at the cost of legal sanctions, to undertake torture.

3.1 The Basic Model

At time 1, an individual or group of individuals in a large population initiates a terrorist action with probability a, and the agency apprehends an individual. If a terrorist action is not initiated, the apprehended person is necessarily innocent. If a terrorist action is initiated, there is a probability b that the apprehended person is guilty. Thus, $\alpha := ab$ is the

probability that the apprehended person has guilty knowledge, and $1-\alpha$ is the probability that the person has no knowledge. Because we are interested in the logic of the ticking bomb scenario, we set $b = 1$, and therefore $\alpha = a$. We are assuming, in other words, that we are in the case in which the benefits from effective torture would be the largest.[12]

At time 2, not knowing whether a terrorist action was initiated, the agency chooses effort $x \geq 0$ to stop a terrorist action by means other than torture. The effort might involve various forms of intelligence gathering and security checks. The cost of x is $c(x)$, where $c' > 0$ and $c'' > 0$. At time 3, if a terrorist action was initiated, Nature chooses whether the agency's effort stops the terrorist action. The probability that the agency stops the terrorist action through its effort x is $\varphi(x)$, where $\varphi' > 0$ and $\varphi'' < 0$.

If the agency does not stop a terrorist action, at time 4, Nature chooses the agency's evidence ε about whether the apprehended individual initiated a terrorist action. The evidence can be high, ε_H, or low, ε_L. The probability of high evidence is $q_1(x)$ if a terrorist action was initiated and $q_2(x)$ if a terrorist action was not initiated. Put another way, if we think of ε_L as evidence indicating innocence of the apprehended person, then $1-q_1(x)$ is the probability of false exculpatory evidence and $1-q_2(x)$ is the probability of accurate exculpatory evidence. We assume that $q_1(x) \geq q_2(x)$ for any $x > 0$ and that $q_1(0) = q_2(0) = y \in (0,1)$, $q_1' > 0$, $q_1'' < 0$, $q_2' < 0$, and $q_2'' > 0$.[13]

At time 5, if torture is legal, the agency chooses whether or not to torture the apprehended individual (T or $\neg T$). We initially assume that the agency never uses torture illegally. If torture is not used and a terrorist action was initiated, the terrorist action succeeds, causing social damages D. If a terrorist action was initiated and torture is used, Nature chooses whether the torture succeeds in stopping the action. With probability θ, the agency extracts

[12]We have checked that this assumption does not change any of the qualitative results.
[13]As specified, if the agency increases x, it obtains a more accurate signal, but the agency still gets a signal even if it chooses $x = 0$.

the critical information from the guilty individual and stops the terrorist action, and with probability $1-\theta$, torture is ineffective and the terrorist action succeeds.[14] If torture is used and no terrorist action was initiated, society incurs damages t from torturing an innocent individual.

We assume that the agency only internalizes a fraction, δ, of social damages, D, and does not internalize the costs of torturing innocent people, t, which is part of the agency problem. This assumption motivates possible constraints on the agency's behavior. If the agency is not as concerned as society about protecting safety or individual rights, society may want to prevent the agency from using torture rather than leave the decision to use torture at the agency's discretion.[15]

If a terrorist action is initiated but the agency stops it (either through torture or other means) or if a terrorist action is not initiated, then the agency's payoff is $-c(x)$. If a terrorist action is initiated and the agency does not stop the terrorist action (because the agency cannot or does not use torture or because torture is unsuccessful), then the agency's payoff is $-\delta D - c(x)$.[16]

We compare outcomes when

1. torture is illegal whether the evidence is low or high, regime B,

2. torture is legal only when evidence is high, regime H, and

[14]As specified, x increases the probability q_1 that the agency has incriminating evidence if the apprehended individual is guilty but does not affect the probability θ that torturing the guilty individual yields the information necessary to stop an attack. We could also make θ an increasing function of x. However, this would not affect the main qualitative results of the paper, although it would add an additional source of complementarity between the agency's use of torture and its other efforts (see footnote 17 below).

[15]We think of this as an extreme version of a reduced form of the difference between agency incentives and society's preferences. For a general analysis of optimal agency discretion when agency objectives are different from those of society, see Shavell (2007).

[16]The agency's payoff might also be affected by whether an attack did not occur because the agency was able to stop it or rather because it was not initiated. However, this generalization would not affect the comparisons of agency effort levels across the regimes that we analyze (see footnote 17 below).

3. torture is legal whether the evidence is high or low, regime LH,

all assuming that the agency obeys directives.

The three regimes have effects on agency behavior, measured by the efforts put into non-torture activities, and on welfare, measured by public safety, the probability of torturing the innocent, and the cost of efforts.

3.2 Agency Behavior

Let $\mathbb{D}(x) = \alpha(1-\varphi(x))[-\delta D]$ denote expected damages under regime B.

1. Under regime B, the agency does not use torture whether it has low or high evidence, strategy $(\neg T, \neg T)$. Thus, its optimal action, x_B^*, solves

$$\max_{x\geq 0} EU^B_{Agency}(x|\neg T, \neg T) = \mathbb{D}(x) - c(x). \tag{1}$$

2. Under regime H, the agency uses tortures only when it has high evidence, strategy $(\neg T, T)$. Thus, its optimal action, x_H^*, solves

$$\max_{x\geq 0} EU^H_{Agency}(x|\neg T, T) = \psi(x)\mathbb{D}(x) - c(x), \tag{2}$$

where $\psi(x) = (1 - q_1(x)\theta)$ is the probability that strong evidence and torture fail.

3. Under regime LH, the agency uses torture whether it has low or high evidence, strategy (T, T). Thus, its optimal action, x_{LH}^*, solves

$$\max_{x\geq 0} EU^{LH}_{Agency}(x|T, T) = (1-\theta)\mathbb{D}(x) - c(x), \tag{3}$$

where $(1-\theta)$ is the probability that torture of a guilty person fails.

We now characterize conditions under which a total ban elicits higher effort than a partial ban and show that having no ban on torture unambiguously reduces other agency effort. (Proofs are in the appendix.)

Proposition 1 *The solutions x_B^*, x_H^*, and x_{LH}^* satisfy*
(a) $x_B^* > x_H^*$ *iff* $\mathbb{D}'(x_H^*) > [\psi(x_H^*)\mathbb{D}(x_H^*)]'$, *and*
(b) $\min(x_B^*, x_H^*) > x_{LH}^*$.

Proposition 1(a) shows that partial bans on torture have two effects on agency efforts to counter terrorism by means other than torture, a decommitment effect and a complementarity effect. Intuitively, allowing the agency to torture when evidence of terrorist action is high reduces the incentives to avoid these situations, a decommitment effect that reduces other efforts. On the other hand, if other efforts raise the chances of having high enough evidence to torture and torture is effective, then torture and other efforts are complementary.

We see these two effects by noting that $\mathbb{D}'(x_H^*) > [\psi(x_H^*)\mathbb{D}(x_H^*)]'$ if and only if

$$\alpha\theta\delta D\left[q_1(x_H^*)\varphi'(x_H^*) - q_1'(x_H^*)(1-\varphi(x_H^*))\right] > 0. \tag{4}$$

The term $q_1(x_H^*)\varphi'(x_H^*)$ is positive and measures the decommitment effect of agency torture on other agency efforts. To the extent that other agency efforts increase the probability of stopping an attack without using torture ($\varphi' > 0$) when they generate enough evidence to use torture (q_1), allowing torture reduces other agency efforts. The term $-q_1'(x_H^*)(1-\varphi(x_H^*))$ is negative and measures the complementarity effect of other agency efforts and agency torture. To the extent that other agency efforts increase the probability of having enough evidence to use torture ($q_1' > 0$) when they fail to stop an attack without using torture ($1-\varphi$), allowing torture increases other agency efforts.

If other agency efforts are more effective in stopping an attack without using torture, i.e., φ' and φ are larger, then the decommitment effect of legalizing torture when evidence is high, $q_1(x_H^*)\varphi'(x_H^*)$, is more likely to dominate its complementarity effect, $-q_1'(x_H^*)(1-\varphi(x_H^*))$, and thus legalizing torture when evidence is high is more likely to reduce other agency efforts. Equation (4) also reveals that, if legalizing torture when evidence is high reduces

other agency efforts, then it reduces other agency efforts by a greater extent if the prior probability of a terrorist attack and the damages resulting from an attack are greater, i.e., α and D are greater. The effect of legalizing torture in high evidence cases on other agency efforts is larger if the threat of an attack is greater.[17]

According to Proposition 1(b), allowing torture even when evidence is low unambiguously reduces the agency's other efforts. Allowing torture even when evidence is low does not increase the chances of having enough evidence to warrant the use of torture in the event that other efforts fail, since the agency always has enough evidence to warrant torture if torture is allowed even when evidence is low. Thus, allowing torture even when evidence is low has no complementarity effect and only a decommitment effect on other efforts and therefore unambiguously reduces other efforts.

3.3 Welfare Effects

We first analyze two central components of welfare, the probability of being safe from a terrorist action and the probability of torturing the innocent.

1. If the agency does not torture, then the probability of safety and the probability of torturing the innocent are, respectively,

$$S_B^* = 1 - \alpha(1 - \varphi(x_B^*)) \text{ and } Q_B^* = 0. \tag{5}$$

[17]If, in addition to making q_1 an increasing and concave function of x, we were to also make θ an increasing and concave function of x, then the condition in (4) would become

$$\alpha\delta D\left[\theta(x_H^*)\{q_1(x_H^*)\varphi'(x_H^*) - q_1'(x_H^*)(1-\varphi(x_H^*))\} - \theta'(x_H^*)q_1(x_H^*)(1-\varphi(x_H^*))\right] > 0.$$

The complementarity effect would then have the additional term $-\theta'(x_H^*)q_1(x_H^*)(1-\varphi(x_H^*))$.

If the agency were to receive an additional payoff F when an attack did not occur because the agency was able to stop it (and no additional payoff when an attack did not occur because it was not initiated), the condition in (4) would become

$$\alpha\theta(\delta D + F)\left[q_1(x_H^*)\varphi'(x_H^*) - q_1'(x_H^*)(1-\varphi(x_H^*))\right] > 0.$$

This generalization would not affect the comparison of agency effort levels across regimes B and H.

2. If the agency tortures only when it has high evidence, then the probabilities are

$$S^*_H = 1 - \alpha(1 - \varphi(x^*_H))\psi(x^*_H) \text{ and } Q^*_H = (1 - \alpha)q_2(x^*_H). \tag{6}$$

3. If the agency tortures on the basis of any evidence at all, then the probabilities are

$$S^*_{LH} = 1 - \alpha(1 - \varphi(x^*_{LH}))(1 - \theta) \text{ and } Q^*_{LH} = (1 - \alpha). \tag{7}$$

The three safety probabilities are strictly increasing in agency effort, and the three probabilities of torturing the innocent are either flat at 0 or strictly decreasing in agency effort. Changes in agency behavior cannot change the part of the welfare ranking due to torturing the innocent–restricting the circumstances under which torture is allowed unambiguously reduces the likelihood of torturing the innocent.

Corollary 1.1 *The probabilities of torturing the innocent,* Q^*_B*,* Q^*_H*, and* Q^*_{LH}*, satisfy* $0 = Q^*_B < Q^*_H < Q^*_{LH}$*.*

By contrast, changes in agency behavior can change the part of the welfare ranking due to safety. Legalizing torture has a direct effect on security–it stops terrorist attacks when they have been initiated and a torture-susceptible guilty person is available. By altering agency behavior, it also has an indirect effect on security. The effect on safety depends on which of these effects dominates.

Corollary 1.2 *The safety probabilities,* S^*_B*,* S^*_H*, and* S^*_{LH}*, satisfy*

(a) $S^*_B > S^*_{LH}$ *iff* $[\varphi(x^*_B) - \varphi(x^*_{LH})] > \theta[1 - \varphi(x^*_{LH})]$,
(b) $S^*_H > S^*_{LH}$ *iff* $[\varphi(x^*_H) - \varphi(x^*_{LH})] > \theta[1 - \varphi(x^*_{LH}) + q_1(x^*_H)(1 - \varphi(x^*_H)]$, *and*
(c) $S^*_B > S^*_H$ *iff* $[\varphi(x^*_B) - \varphi(x^*_H)] > \theta q_1(x^*_H)[1 - \varphi(x^*_H)]$.

In parts (a) and (b) of Corollary 1.2, the terms on the left-hand side of the inequalities are unambiguously positive since $\min(x^*_B, x^*_H) > x^*_{LH}$ by Proposition 1(b). If $\theta = 0$ in

(a) and (b), then $S^*_{LH} < \min\{S^*_H, S^*_B\}$. Allowing torture whether evidence is high or low unambiguously reduces security if torture is highly ineffective. If $\theta = 1$ in (a) and (b), then $S^*_{LH} > \max\{S^*_H, S^*_B\}$. If torture is nearly foolproof, then it is the easiest way to guarantee security, in which case a tradeoff exists between safety and torture of the innocent.

In part (c) of Corollary 1.2, the left-hand side is positive if the decommitment effect of legalizing torture when evidence is high dominates its complementarity effect as identified in Proposition 1(a), so that $x^*_B > x^*_H$. If x^*_B is sufficiently greater than x^*_H, then $S^*_B > S^*_H$. As noted above, legalizing torture when evidence is high is more likely to reduce agency effort if agency effort is more effective in stopping an attack without using torture. Moreover, if legalizing torture when evidence is high reduces agency effort, it reduces agency effort by a greater extent if the attack threat is greater. Thus, if agency effort is sufficiently effective in stopping an attack without using torture and the threat of an attack is sufficiently great, legalizing torture when evidence is high reduces safety.

If the availability of torture sufficiently reduces other agency efforts, it reduces safety and increases torture of the innocent; otherwise, it increases safety and torture of the innocent.

The full welfare levels under each of the regimes are

$$W^*_B = -(1 - S^*_B)D - c(x^*_B),\ W^*_H = -(1 - S^*_H)D - Q^*_H t - c(x^*_H), \text{ and} \tag{8}$$

$$W^*_{LH} = -(1 - S^*_{LH})D - Q^*_{LH} t - c(x^*_{LH}). \tag{9}$$

The agency's payoffs under each of the regimes are the same as society's payoffs except with D replaced by δD and with $t = 0$. The next corollary compares welfare across regimes.

Corollary 1.3 *The welfare levels, W^*_B, W^*_H, and W^*_{LH}, satisfy*

(a) $W^*_B > W^*_{LH}$ *iff* $[S^*_B - S^*_{LH}]D + Q^*_{LH} t > [c(x^*_B) - c(x^*_{LH})]$,
(b) $W^*_H > W^*_{LH}$ *iff* $[S^*_H - S^*_{LH}]D + [Q^*_{LH} - Q^*_H]t > [c(x^*_H) - c(x^*_{LH})]$, *and*
(c) $W^*_B > W^*_H$ *iff* $[S^*_B - S^*_H]D + Q^*_H t > [c(x^*_B) - c(x^*_H)]$.

If the availability of torture reduces safety, both sides of the inequalities in Corollary 1.3 are unambiguously positive. In this case, the availability of torture reduces welfare if the social damages from a terrorist attack, D, or the social damages from torturing the innocent, t, are sufficiently high. However, if the extent δ to which the agency internalizes D is not too high, then the availability of torture increases the agency's payoff although it reduces welfare. In such a case, the agency has incentives to disobey directives on torture, which leads us to the enforcement problem.

4. The Enforcement Problem

Our analysis of the agency problem has focused on the choice of non-torture efforts assuming that the agency obeys directives on torture. We now suppose that the agency faces a penalty, p, if it tortures when not allowed to.[18] In this context, we examine what happens when

1. torture is banned by penalties, regime $\mathcal{B}(p,p)$ with penalties $(p_L, p_H) = (p, p)$,
2. torture is regulated by a warrant system, regime $\mathcal{W}(p,0)$, in which torture does not carry a penalty if evidence is high, i.e., the penalties are $(p_L, p_H) = (p, 0)$, and
3. torture is not penalized, regime $\mathcal{LH}$ with penalties $(p_L, p_H) = (0, 0)$.

Regime B of the previous section achieves a complete torture ban because the agency obeys directives, while regime $\mathcal{B}(p,p)$ penalizes any torture, and only achieves the ban if the penalty is sufficiently high. The comparison between the regimes H and $\mathcal{W}(p,0)$ is similar. Though we analyze this regime as a system of torture warrants, it is formally identical to nonprosecution of torture in high evidence cases, however achieved.

[18]The penalty might correspond to the utility cost of the possibility of being prosecuted within the domestic judiciary system or being declared a war criminal by international courts after choosing to use torture. *Ex post* enforcement creates a "liability rule" against torture. For an economic analysis of liability rules in the protection of individual rights, see Kontorovich (2004) and Kaplow and Shavell (1996).

We first gather results that allow us to study agency behavior under the three regimes. With these in place, we then analyze, within the context of the model, the Dershowitz arguments for reductions in the frequency of torture arising from a torture warrant system. We then turn to the existence of slippery slopes. These are situations in which legalizing torture under some circumstances leads to a yet larger increase in the set of circumstances under which torture occurs.

It is the intermediate values of p that lead to the analysis of the Dershowitz arguments and/or the slippery slopes. Large enough values of p discourage all illegal torture. For regime $\mathcal{B}$, this means that $(\neg T, \neg T)$ is optimal for large p, and for regime $\mathcal{W}$, $(\neg T, T)$ is optimal. By contrast, for smaller values of p, (T, T) is optimal for all three regimes.

4.1 Agency Behavior

Let $P(\varepsilon_L|x)$ and $P(\varepsilon_H|x)$ be the likelihoods of a low and a high evidence suspect, and let $\beta_L(x)$ and $\beta_H(x)$ be the likelihoods of a low and a high evidence suspect being guilty, respectively. From Bayes' rule, we have

$$\beta_L(x) = \frac{\alpha(1-\varphi(x))(1-q_1(x))}{P(\varepsilon_L|x)} \text{ and} \tag{10}$$

$$\beta_H(x) = \frac{\alpha(1-\varphi(x))q_1(x)}{P(\varepsilon_H|x)}, \text{ where} \tag{11}$$

$$P(\varepsilon_L|x) = \alpha(1-\varphi(x))(1-q_1(x)) + (1-\alpha)(1-q_2(x)) \text{ and} \tag{12}$$

$$P(\varepsilon_H|x) = \alpha(1-\varphi(x))q_1(x) + (1-\alpha)q_2(x). \tag{13}$$

Note that $P(\varepsilon_L|x)$ and $P(\varepsilon_H|x)$ do not sum to one, since there is the probability of no suspect anymore because x has caused the attack to be averted.

We first gather three useful observations.

Lemma 1 *For all x,*
(a) $\beta_L(x) < \beta_H(x)$ *and* $\beta'_L(x) < 0$,
(b) *the agency's optimal choice is* T *at* ε_L *if* $\beta_L(x) > \frac{p_L}{\theta\delta D}$ *and* T *at* ε_H *if* $\beta_H(x) > \frac{p_H}{\theta\delta D}$, *and*
(c) $[P(\varepsilon_L|x) + P(\varepsilon_H|x)]' < 0$, *which implies that at least one of the derivatives,* $[P(\varepsilon_L|x)]'$ *and* $[P(\varepsilon_H|x)]'$, *is negative.*

β_L is decreasing in x because increases in x reduce the probability of a terrorist action eluding the agency's non-torture efforts, reduce the probability of false exculpatory evidence, and increase the probability of valid exculpatory evidence. It is perhaps intuitive that β_H should be increasing in x because increases in x reduce the probability of false evidence of guilt. However, since increases in x reduce the probability of a terrorist action eluding the agency's non-torture efforts, the likelihood of true evidence of guilt may decrease enough with x to offset this effect.

The agency's behavior under the different regimes can be understood in two parts: (A) for a given torture policy, finding the agency's optimal effort, and (B) comparing the payoffs to the different torture policies given optimal agency effort levels. From the first two parts of Lemma 1, we know that the agency's torture policy is either (T, T), $(\neg T, T)$, or $(\neg T, \neg T)$ under regime $\mathcal{B}$, it is either (T, T) or $(\neg T, T)$ under regime $\mathcal{W}$, and it is always (T, T) under regime $\mathcal{LH}$.

For regime $\mathcal{B}$, let $f_{\mathcal{B}}^{T,T}(p)$, $f_{\mathcal{B}}^{\neg T,T}(p)$, and $f_{\mathcal{B}}^{\neg T,\neg T}(p)$ be the agency's value functions at the optimal effort levels for a given p when the agency's policy is (T, T), $(\neg T, T)$, and $(\neg T, \neg T)$, respectively. Then, under regime $\mathcal{B}$, agency behavior is the solution to the problem

$$V_{\mathcal{B}}(p) = \max\{f_{\mathcal{B}}^{T,T}(p), f_{\mathcal{B}}^{\neg T,T}(p), f_{\mathcal{B}}^{\neg T,\neg T}(p)\}, \text{ where} \tag{14}$$

$$f_{\mathcal{B}}^{T,T}(p) = \max_{x\geq 0}[(1-\theta)\mathbb{D}(x) - c(x)] - p\left\{P(\varepsilon_L|x) + P(\varepsilon_H|x)\right\}, \tag{15}$$

$$f_{\mathcal{B}}^{\neg T,T}(p) = \max_{x\geq 0}[\psi(x)\mathbb{D}(x) - c(x)] - p\left\{0 + P(\varepsilon_H|x)\right\}, \text{ and} \tag{16}$$

$$f_{\mathcal{B}}^{\neg T,\neg T}(p) = \max_{x\geq 0}[\mathbb{D}(x) - c(x)] - p\left\{0 + 0\right\}. \tag{17}$$

Similarly, for regime W, let $f_{\mathcal{W}}^{T,T}(p)$ and $f_{\mathcal{W}}^{\neg T,T}(p)$ be the agency's value functions at the optimal efforts for a given p when the agency's policy is (T,T) and $(\neg T,T)$, respectively. Agency behavior is the solution to the problem

$$V_{\mathcal{W}}(p) = \max\{f_{\mathcal{W}}^{T,T}(p), f_{\mathcal{W}}^{\neg T,T}(p)\}, \text{ where} \tag{18}$$

$$f_{\mathcal{W}}^{T,T}(p) = \max_{x\geq 0}[(1-\theta)\mathbb{D}(x) - c(x)] - p\left\{P(\varepsilon_L|x) + 0\right\}, \text{ and} \tag{19}$$

$$f_{\mathcal{W}}^{\neg T,T}(p) = \max_{x\geq 0}[\psi(x)\mathbb{D}(x) - c(x)] - p\left\{0+0\right\}. \tag{20}$$

Under regime $\mathcal{LH}$, p does not play a role, and the maximization problem is

$$f_{\mathcal{LH}}^{T,T} = \max_{x\geq 0}[(1-\theta)\mathbb{D}(x) - c(x)] - p\left\{0+0\right\}. \tag{21}$$

When the penalties p are equal to 0, all of the regimes are equivalent, and equivalent to the regime LH of the previous section. This implies that

$$V_{\mathcal{B}}(0) = V_{\mathcal{W}}(0) = f_{\mathcal{B}}^{T,T}(0) = f_{\mathcal{W}}^{T,T}(0) = f_{\mathcal{LH}}^{T,T}. \tag{22}$$

Note that under both regimes $\mathcal{B}$ and $\mathcal{W}$, as the agency tortures in fewer circumstances, i.e., as we move down through the three cases in (15), (16), and (17) under regime $\mathcal{B}$, and down through the two cases in (19) and (20) under regime $\mathcal{W}$, for all fixed values of x, the terms that multiply the penalties p decrease. For example, under regime $\mathcal{B}$, we have the terms $\{P(\varepsilon_L|x) + P(\varepsilon_H|x)\}$, then $\{0 + P(\varepsilon_H|x)\}$, and then $\{0+0\}$. This suggests that the value functions in the two regimes have a "single-crossing from above in p" property. This turns out to be true but is somewhat more subtle because we are evaluating the penalty terms at different values of x. The following lemma addresses this issue and gives additional useful properties of the value functions.

Lemma 2 *Under regimes $\mathcal{B}$ and $\mathcal{W}$, the value function for any policy that tortures is strictly decreasing in p, convex in p, and crosses the value function for a policy that tortures less exactly once, from above, as p increases.*

Define $\underline{p_{\mathcal{B}}}$, $\overline{p_{\mathcal{B}}}$, and $p_{\mathcal{W}}$ to be the penalty levels at which we have the crossings,

$$f_{\mathcal{B}}^{T,T}(\underline{p_{\mathcal{B}}}) = f_{\mathcal{B}}^{\neg T,T}(\underline{p_{\mathcal{B}}}),\ f_{\mathcal{B}}^{\neg T,T}(\overline{p_{\mathcal{B}}}) = f_{\mathcal{B}}^{\neg T,\neg T}(\overline{p_{\mathcal{B}}}),\text{ and} \tag{23}$$

$$f_{\mathcal{W}}^{T,T}(p_{\mathcal{W}}) = f_{\mathcal{W}}^{\neg T,T}(p_{\mathcal{W}}). \tag{24}$$

4.2 The Dershowitz Argument

As noted above, the core of the Dershowitz (2002) argument is that torture happens although it is illegal and that an enforced system of judicial warrants could bring this under control, resulting in less torture. In our model, Dershowitz's case corresponds to regime $\mathcal{B}(p^\circ, p^\circ)$ for a p° at which $V_{\mathcal{B}}(p^\circ) = f_{\mathcal{B}}^{\neg T,T}(p^\circ)$, i.e., the agency's torture policy is $(\neg T, T)$. The appropriate comparison is the change of regime to the regime $\mathcal{W}(p^\circ, 0)$. From Lemma 1(b), the agency's choice of torture policy in regime $\mathcal{W}(p^\circ, 0)$ is either $(\neg T, T)$, the case we analyze now, or (T, T), which we analyze in Proposition 4 as part of our discussion of slippery slopes.

We first provide conditions under which the agency's optimal torture policy is $(\neg T, T)$ under both regimes $\mathcal{B}$ and $\mathcal{W}$.

Proposition 2 *The agency's optimal torture policy is $(\neg T, T)$ under both regimes $\mathcal{B}(p^\circ, p^\circ)$ and $\mathcal{W}(p^\circ, 0)$ if $\underline{p_{\mathcal{B}}} < \overline{p_{\mathcal{B}}}$, $p_{\mathcal{W}} < \overline{p_{\mathcal{B}}}$, and $p^\circ \in (\max\{p_{\mathcal{W}}, \underline{p_{\mathcal{B}}}\}, \overline{p_{\mathcal{B}}})$.*

We now compare agency effort, safety, and the frequency of torture of the innocent under regimes $\mathcal{B}$ and $\mathcal{W}$ in this case.

Proposition 3 *Suppose the agency's optimal torture policy is $(\neg T, T)$ under both regimes $\mathcal{B}(p^\circ, p^\circ)$ and $\mathcal{W}(p^\circ, 0)$.*

(a) *If $[P(\varepsilon_H|x)]' = \alpha[q_1'(x)(1-\varphi(x)) - \varphi'(x)q_1(x)] + (1-\alpha)q_2'(x) < 0$ for all x, then regime $\mathcal{W}(p^\circ, 0)$ has lower effort and therefore lower safety and more frequent torture of the innocent.*

(b) *If $[P(\varepsilon_H|x)]' = \alpha[q_1'(x)(1-\varphi(x)) - \varphi'(x)q_1(x)] + (1-\alpha)q_2'(x) > 0$ for all x, then regime $\mathcal{W}(p^\circ, 0)$ has higher effort and therefore higher safety and less frequent torture of the innocent.*

The derivative $[P(\varepsilon_H|x)]'$ plays a crucial role. Its sign determines whether the relevant penalty for illegal torture, $-p^\circ\{0 + P(\varepsilon_H|x)\}$ in (16), is increasing or decreasing in agency effort. The absence of this penalty is the only difference between (16) and (20), and (20) describes agency effort choice in regime $\mathcal{W}$.

Note that $P(\varepsilon_H|x)$, which is given in (13), has two parts. The second part, $(1-\alpha)q_2(x)$, is the likelihood of the evidence accusing an innocent person of having knowledge of a terrorist action, and this likelihood is strictly decreasing in x. The first part, $\alpha(1-\varphi(x))q_1(x)$, is the likelihood of a terrorist action not being stopped by agency efforts, $\alpha(1-\varphi(x))$, which is decreasing in x, times the probability $q_1(x)$ that the evidence accuses a person with guilty knowledge, which is increasing in x. If this product is decreasing, then $[P(\varepsilon_H|x)]' < 0$. If the product is increasing, we may still have a negative derivative if e.g. $\alpha(1-\varphi(x))$ is close to 0, that is, if the agency's non-torture efforts are effective.

Explicit consideration of the derivative is instructive. We have

$$[P(\varepsilon_H|x)]' = \alpha[q_1'(x)(1-\varphi(x)) - \varphi'(x)q_1(x)] + (1-\alpha)q_2'(x). \tag{25}$$

Thus, the sign of $[P(\varepsilon_H|x)]'$ depends on the relative sizes of the three terms $\alpha q_1'(x)(1-\varphi(x))$, $-\alpha\varphi'(x)q_1(x)$, and $(1-\alpha)q_2'(x)$.

Moving from $\mathcal{B}$ to $\mathcal{W}$ tends to reduce agency effort x because it eliminates the agency's cost of using torture if the individual is guilty and the agency's evidence turns out to be high, which reduces the agency's commitment not to use torture. This is the decommitment effect, which is captured by the negative term $-\varphi'(x)q_1(x)$.[19] Moving from $\mathcal{B}$ to $\mathcal{W}$ also tends to increase x because increasing x increases the probability that the agency has high evidence if the individual is guilty, in which case the agency can use torture without punishment. This is

[19]Although the agency will torture a suspect if and only if it has high evidence regardless of whether it will be punished for this (regime $\mathcal{B}$) or not (regime $\mathcal{W}$), if it will be punished for doing so (regime $\mathcal{B}$), then it has a greater incentive to invest in preventative effort x in order to increase the probability of stopping the terrorist attack without needing to torture and thus subjecting itself to punishment.

the complementarity effect, which is captured by the positive term $q_1'(x)(1-\varphi(x))$. Moving from $\mathcal{B}$ to $\mathcal{W}$ also tends to reduce x because reducing x increases the probability that the agency has high evidence if the individual is innocent, in which case the agency can escape punishment for using torture on an innocent individual. This is the decomplementarity effect, which is captured by the negative term $(1-\alpha)q_2'(x)$.

If the complementarity effect dominates both the decommitment and complementarity effects, then moving from $\mathcal{B}$ to $\mathcal{W}$ increases agency effort x, and thereby actually increases safety and reduces the probability of torturing the innocent. However, if the decommitment and complementarity effects together dominate the complementarity effect, then moving from $\mathcal{B}$ to $\mathcal{W}$ reduces agency effort x, and thereby reduces safety and increases the probability of torturing the innocent.

4.3 Slippery Slopes

We have a slippery slope if legalizing torture in dire circumstances increases the set of circumstances in which torture occurs. There are two ways in which this could arise in the model: $(\neg T, T)$ is the agency's choice under regime $\mathcal{B}(p^\circ, p^\circ)$ while (T, T) is the choice under regime $\mathcal{W}(p^\circ, 0)$; and $(\neg T, \neg T)$ is the agency's choice under regime $\mathcal{B}(p^\circ, p^\circ)$ while (T, T) is the choice under regime $\mathcal{W}(p^\circ, 0)$. In either case, switching to a torture warrant system lowers agency effort and increases the frequency of torture of the innocent.

Proposition 4 *If p° is such that either*

(a) *the agency's optimal torture policy is $(\neg T, \neg T)$ under regime $\mathcal{B}(p^\circ, p^\circ)$ and is (T, T) under regime $\mathcal{W}(p^\circ, 0)$, or*
(b) *the agency's optimal torture policy is $(\neg T, T)$ under regime $\mathcal{B}(p^\circ, p^\circ)$ and is (T, T) under regime $\mathcal{W}(p^\circ, 0)$,*

then switching from regime $\mathcal{B}(p^\circ, p^\circ)$ to regime $\mathcal{W}(p^\circ, 0)$ (the torture warrant system) reduces agency effort and increases the frequency of torture of the innocent.

Intuitively, if the torture policy changes in either one of the ways indicated, then this is direct evidence of lower agency effort. More formally, from Lemma 1(a), the conditional probability of a low evidence suspect being guilty, $\beta_L(x)$, satisfies $\beta'_L(x) < 0$ for all x. For the regime $\mathcal{W}(p^\circ, 0)$, Lemma 1(b) implies that $\beta_L(x) > p^\circ/\theta\delta D$ leads to torture when evidence is low. Since $\beta'_L(x) < 0$, only a lower x can lead the agency to torture when it sees low evidence. A lower x in turn increases the frequency of torture of the innocent.

Whether or not moving down either type of slippery slope translates to lower safety depends, as in Corollary 1.2, on the effectiveness of torture. If torture is sufficiently effective, i.e., if θ is sufficiently close to 1, then torturing more increases safety.

However, the tradeoffs between safety and torture of the innocent may depend on the type of slippery slope. In a type (a) slippery slope, moving from regime $\mathcal{B}$ to regime $\mathcal{W}$ leads the agency to use torture when evidence is low, which may greatly increase torture of the innocent, but it also leads the agency to use torture when evidence is high, which may greatly increase security if torture is very effective. In a type (b) slippery slope, moving from $\mathcal{B}$ to $\mathcal{W}$ also leads the agency to use torture when evidence is low, but it does not change the agency's torture choice when evidence is high, since the agency is already using torture when evidence is high under $\mathcal{B}$. Therefore it may increase security only minimally even if torture is very effective but may still extensively increase torture of the innocent.

Whether or not each type of slippery slope arises depends on the levels of the crossing points of the value functions for regime $\mathcal{B}$, $\underline{p_{\mathcal{B}}}$ and $\overline{p_{\mathcal{B}}}$, defined in (23), and the crossing point of the value functions for regime $\mathcal{W}$, $p_{\mathcal{W}}$, defined in (24).

Proposition 5 *Suppose* $\underline{p_{\mathcal{B}}} < \overline{p_{\mathcal{B}}}$.

(a) *The agency's optimal torture policy is* $(\neg T, \neg T)$ *under regime* $\mathcal{B}(p^\circ, p^\circ)$ *and is* (T, T) *under regime* $\mathcal{W}(p^\circ, 0)$ *if* $\overline{p_{\mathcal{B}}} < p_{\mathcal{W}}$ *and* $p^\circ \in (\overline{p_{\mathcal{B}}},\, p_{\mathcal{W}})$.

(b) *The agency's optimal torture policy is* $(\neg T, T)$ *under regime* $\mathcal{B}(p^\circ, p^\circ)$ *and is* (T, T) *under regime* $\mathcal{W}(p^\circ, 0)$ *if* $\underline{p_{\mathcal{B}}} < p_{\mathcal{W}}$ *and* $p^\circ \in (\underline{p_{\mathcal{B}}},\, \min\{\overline{p_{\mathcal{B}}},\, p_{\mathcal{W}}\})$.

The crossing point conditions in Proposition 5(a) and (b) are more easily satisfied if the value functions for policies involving torture in regime $\mathcal{B}$ are steeper while the value function for the policy involving torture in regime $\mathcal{W}$ is shallower. The value functions in $\mathcal{B}$ are relatively steeper if the level of $P(\varepsilon_H|x)$ is higher, because this term is the difference between the penalties across $\mathcal{B}$ and $\mathcal{W}$. If the probability of having high evidence suspects is higher, the agency is more tempted to rely on torture as a counterterrorism tool, and moving to $\mathcal{W}$ is more likely to lower agency effort and increase torture of the innocent. Moreover, if $P(\varepsilon_H|x)$ does not vary much with x, then the value functions in $\mathcal{B}$ are less convex, meaning that they stay steeper for a larger range of penalties. If the probability of having a high evidence suspect is less sensitive to effort, then there is less loss to lowering effort if torture is used more, which also makes a slippery slope more likely to arise.

The slippery slopes in (a) and (b) both arise from lower effort reducing the quality of exculpatory evidence. Intuitively, moving from regime $\mathcal{B}$ to regime $\mathcal{W}$ eliminates the agency's penalty for using torture in high evidence cases. This reduces the agency's commitment to non-torture efforts, whether or not the agency was already using torture in high evidence cases under $\mathcal{B}$. A reduction in non-torture efforts reduces the quality of the agency's evidence, which increases the agency's incentives to adopt a strategy that uses torture even when evidence is low. Adoption of such a strategy further reduces agency efforts, further reinforcing the agency's incentives to use torture even when evidence is low.

5. Endogenous Terrorism

Our analysis so far has assumed that the probability of a terrorist attack, α, is the same regardless of the legal regime. We now endogenize the probability of terrorist attack and show that legalizing torture may increase the probability of attack.

5.1 Agency and Terrorist Behavior

We first consider a model that is identical to the basic model in Section 3 except that individuals have a choice of whether to initiate a terrorist action. Individuals now differ according to their benefit from initiating a terrorist attack, w. At time 1, Nature chooses each individual's w according to the cumulative density function $F(w)$, which is assumed to be differentiable and have inverse function $F^{-1}(\cdot)$. The associated probability density function is denoted by $f(w)$. At time 2, individuals each choose whether or not to become a terrorist and initiate terrorist action (TA or $\neg TA$). The cost of initiating terrorist action is denoted by c.

The rest of the model's timing is the same as in the basic model. At time 3, without knowing whether a terrorist action was initiated, the agency chooses effort $x \geq 0$ to stop a terrorist action by means other than torture. At time 4, if a terrorist action was initiated, the agency stops it with probability $\varphi(x)$, where $\varphi' > 0$ and $\varphi'' < 0$. If the agency does not stop a terrorist action, at time 5, Nature chooses whether the agency receives high or low evidence about whether an individual initiated a terrorist action. The probability of high evidence is $q_1(x)$ if a terrorist action was initiated and $q_2(x)$ if a terrorist action was not initiated, where $q_1(x) \geq q_2(x)$. At time 6, if the evidence is high and torture is legal, the agency chooses whether or not to use torture (T or $\neg T$). If torture is not used and a terrorist action was initiated, the terrorist action succeeds. If a terrorist action was initiated and torture is used, then with probability θ, torture is successful and the terrorist action is stopped, and with probability $1 - \theta$, torture is ineffective and the terrorist action succeeds.

An individual who has chosen to become a terrorist incurs a cost t_G from being tortured, while an individual who is not a terrorist incurs a cost t_I from being tortured.

Denote by σ_τ the probability that the agency uses torture when the evidence is high.

For given x and σ_τ, the expected payoffs to a type w individual of becoming a terrorist and initiating a terrorist action, TA, and from not becoming a terrorist, $\neg TA$, are, respectively,

$$EU_{Individual}(TA) = (1-\varphi(x))(1-q_1(x)\sigma_\tau\theta)w - (1-\varphi(x))q_1(x)\sigma_\tau t_G - c, \quad (26)$$

$$EU_{Individual}(\neg TA) = -q_2(x)\sigma_\tau t_I. \quad (27)$$

A type w individual chooses TA if and only if

$$w \geq w^*(\sigma_\tau, x) \equiv \frac{[c - \sigma_\tau(q_2(x)t_I - (1-\varphi(x))q_1(x)t_G)]}{(1-\varphi(x))(1-q_1(x)\sigma_\tau\theta)}. \quad (28)$$

The fraction of individuals who initiate terrorist action as a function of the agency's probability of using torture is then

$$\alpha(\sigma_\tau, x) = 1 - F(w^*) = 1 - F\left(\frac{[c - \sigma_\tau(q_2(x)t_I - (1-\varphi(x))q_1(x)t_G)]}{(1-\varphi(x))(1-q_1(x)\sigma_\tau\theta)}\right). \quad (29)$$

Consider now the agency's problem. For a given probability of terrorist action α, if the agency observes high evidence and torture is legal when evidence is high, then the agency always uses torture when evidence is high. That is, $\sigma_\tau = 1$ under regime H. The agency's optimal level of effort x_H^* is then the solution to (2) in Section 3.2. Thus, in equilibrium, the probability of terrorist attack is

$$\alpha_H^* = \alpha(\sigma_\tau = 1, x_H^*) = 1 - F(w_H^*), \quad (30)$$

$$\text{where } w_H^* = \frac{c - (q_2(x_H^*)t_I - (1-\varphi(x_H^*))q_1(x_H^*)t_G)}{(1-\varphi(x_H^*))(1-q_1(x_H^*)\theta)}.$$

On the other hand, under regime B, $\sigma_\tau = 0$ (assuming the agency obeys directives). The agency's optimal level of effort x_B^* is then the solution to (1) in Section 3.2. Thus, in equilibrium, the probability of terrorist attack is

$$\alpha_B^* = \alpha(\sigma_\tau = 0, x_B^*) = 1 - F(w_B^*), \quad (31)$$

$$\text{where } w_B^* = \frac{c}{(1-\varphi(x_B^*))}.$$

Comparing the probabilities of attack under regimes B and H, we have

$$\alpha_H^* - \alpha_B^* < 0 \Leftrightarrow 1 - F(w_H^*) > 1 - F(w_B^*) \Leftrightarrow w_B^* - w_H^* > 0, \tag{32}$$

where

$$w_B^* - w_H^* = \ w^*(\sigma_\tau = 0, x_B^*) - w^*(\sigma_\tau = 0, x_H^*) + \ w^*(\sigma_\tau = 0, x_H^*) - w^*(\sigma_\tau = 1, x_H^*). \tag{33}$$

The term $w^*(\sigma_\tau = 0, x_B^*) - w^*(\sigma_\tau = 0, x_H^*)$ shows the indirect effect (through agency effort) of legalizing torture on the probability of attack, and it is positive if and only if $x_H^* < x_B^*$ since $\frac{dw_B}{dx} = \frac{c\varphi'(x)}{(1-\varphi(x))^2} > 0$. From Proposition 1 in Section 3.2, we know that $x_H^* < x_B^*$ if $\mathbb{D}'(x_H^*) > [\psi(x_H^*)\mathbb{D}(x_H^*)]'$, that is, if the decommitment effect of legalizing torture dominates the complementarity effect.

The term $w^*(\sigma_\tau = 0, x_H^*) - w^*(\sigma_\tau = 1, x_H^*)$ shows the direct effect of legalizing torture on the probability of attack, and it is positive if and only if

$$c < \frac{(q_2(x_H^*)t_I - (1 - \varphi(x_H^*))q_1(x_H^*)t_G)}{q_1(x_H^*)\theta}. \tag{34}$$

Therefore, we have the following sufficient conditions for legalizing torture to increase the probability of terrorist attack:

Proposition 6 *$\alpha_H^* > \alpha_B^*$ if*

$$\mathbb{D}'(x_H^*) > [\psi(x_H^*)\mathbb{D}(x_H^*)]' \text{ and } c < \frac{(q_2(x_H^*)t_I - (1 - \varphi(x_H^*))q_1(x_H^*)t_G)}{q_1(x_H^*)\theta}.$$

Intuitively, if the decommitment effect dominates the complementarity effect ($\mathbb{D}'(x_H^*) > [\psi(x_H^*)\mathbb{D}(x_H^*)]'$), then legalizing torture indirectly increases the probability of attack by reducing the agency's non-torture efforts and making the agency sloppier in its other preventive work, which increases the probability that a terrorist action would succeed.

Moreover, if condition (34) is satisfied, then legalizing torture directly increases the probability of attack by increasing the expected payoff of attacking relative to the expected payoff of not attacking for any given effort level by the agency. The condition is more likely to be satisfied if c is lower (i.e., the costs of initiating a terrorist action are lower), $\varphi(x)$ is higher (i.e., non-torture efforts are more effective at preventing attacks in the first place), θ or $q_1(x)$ are lower (i.e., torture is less effective), and t_I is higher relative to t_G (i.e., the costs to innocents of being tortured are higher relative to the costs to terrorists).

It is quite plausible that condition (34) would be satisfied. While torture can certainly work in particular cases, it may not be too highly effective in general (Costanzo, Gerrity, and Lykes, 2007). Evidence about whether suspects are terrorists is rarely perfectly accurate, and government interrogators may not be able to discern whether suspects are lying about what they know or do not know. In several controlled experiments, police officers with interrogation training have been found to be able to detect deception at a level only slightly higher than chance (Garrido, Masip, and Herrero, 2004). When the government is mistaken about suspects, torturing them cannot yield useful information. Other efforts, such as more sophisticated intelligence gathering and careful security screening, may be more effective.

The costs of being tortured may also be higher for innocents than for terrorists. Unlike terrorists, innocents are not likely to have been exposed to torture or to be prepared for it, so the experience may be more traumatizing for them. Moreover, whereas guilty individuals can make the torture stop by providing the relevant information when the information is verifiable, innocent individuals cannot do so since they do not have the relevant information. There is evidence that interrogators become most coercive when questioning innocent suspects, because truthful suspects are more likely to be regarded as resistant and defiant (Kassin, Goldstein, and Savitsky, 2003).

5.2 Torture as a Signal

Legalizing torture may also increase the probability of terrorist attack for another reason. If illegitimate governments are more prone to use torture than legitimate ones, then legalizing torture might be a signal that a government is illegitimate, which might increase the benefits to individuals of attacking the government and facilitate terrorist recruitment. Mulligan, Gil, and Sala-i-Martin (2004) find that widespread torture is significantly more common in nondemocracies than in democracies. Dreher, Gassebner, and Siemers (2007) find that terrorist attacks in a country are significantly positively associated with human rights violations (including torture) in that country. Krueger and Maleckova (2003) and Kreuger (2007) find that the main factor that raises the likelihood that people from a country will participate in terrorism is the suppression of civil liberties in that country. By engaging in torture, a government risks pooling with illegimate governments and inciting a terrorist backlash.

6. Summary and Future Work

We developed a model of counterterrorism to analyze the effects of allowing the government to use torture when evidence of terrorist involvement is high. We first examined the case in which the agency tasked with counterterrorism places a different weight on torture than society does, but in which it follows any directives. In this case, we showed that allowing the agency to use torture in high evidence cases may reduce its efforts to stop terrorism by means other than torture. This effect blunts any gain to safety that may arise through torture, and the net effect may be a reduction in security.

We then extended our analysis to encompass the possibility that there is an enforcement problem and the agency is willing to disobey torture directives at the risk of legal sanction. Our extension brought to light a slippery slope that works through the endogeneity of the

quality of information rather than through utility changes, cost changes, or patterns of bureaucratic behavior. Allowing torture in high evidence cases may reduce the agency's non-torture efforts. The resulting agency deskilling would reduce the quality of exculpatory evidence. If other preventive efforts fail, this may lead to torture even in low evidence cases.

We also extended the analysis to capture the possible effects of allowing torture on incentives to commit terrorist acts. We showed that greater use of torture by the state may increase the probability of terrorist attacks against the state, which may further increase the state's incentives to use torture to stop attacks.

The main arguments that we developed have a simple outline: loosening constraints on torture may induce changes in terrorist and agency behavior that may compromise security and may even reduce the quality of the agency's evidence to such an extent that it motivates the use of torture even in the face of potentially exculpatory evidence. There may be parallels in arguments for essentially complete freedom of the press. For example, accepting the argument that national security would be endangered by publication of the news that armored vehicles do not protect against roadside bombs might reduce the incentives of those charged with the contracting for and building of armored vehicles to do an adequate job in supplying army equipment. This in turn may make future stories of dangerously defective equipment more likely to be true.

While we considered moral hazard problems with respect to torture, we did not consider potential adverse selection problems. The type of individuals who want to serve enforcement agencies when torture is and is not legal might differ. Once torture is allowed in extreme conditions, more sadistic individuals might want to work in the enforcement agency and naturally "extreme conditions" may become less extreme. It would be interesting to examine the implications of a fuller model with adverse selection as well as moral hazard effects.

A Mathematical Appendix

Proof of Proposition 1. For (a), note that the objective functions in (1), (2), and (3) are the sums of strictly concave functions. This means that optimal agency effort under regime B is characterized by (1′) $\mathbb{D}'(x_B^*) = c'(x_B^*)$, and it is characterized by (2′) $[\psi(x_H^*)\mathbb{D}(x_H^*)]' = c'(x_H^*)$ under regime H. Since $c''(x) > 0$, the derivative condition $\mathbb{D}'(x_H^*) > [\psi(x_H^*)\mathbb{D}(x_H^*)]'$ holds iff $x_B^* > x_H^*$.

The first part of (b), $x_B^* > x_{LH}^*$, follows from $(1-\theta) < 1$ and the supermodularity of $h(x,\theta) := (1-\theta)\mathbb{D}(x) - c(x)$. For the second part of (b), note that under regime LH, optimal agency effort is characterized by (3′) $(1-\theta)\mathbb{D}'(x_{LH}^*) = c'(x_{LH}^*)$. Since $c''(x) > 0$, (2′) and (3′) deliver $[\psi(x_H^*)\mathbb{D}(x_H^*)]' > (1-\theta)\mathbb{D}'(x_H^*)$ iff $x_H^* > x_{LH}^*$. Since $\psi(x) = (1 - q_1(x)\theta)$, rearrangement yields $x_H^* > x_{LH}^*$ iff $q_1'(x_H^*)\mathbb{D}(x_H^*) < (1 - q_1(x_H^*))\mathbb{D}'(x_H^*)$, and the left-hand side is negative while the right-hand side is positive. ∎

Proof of Corollary 1.2. The results are rearrangements of the pairwise differences between the three equations, (5), (6), and (7). ∎

Proof of Corollary 1.1. This follows from $0 < q_2(x) < 1$. ∎

Proof of Lemma 1. For the first part of (a), note that $\beta_L(x) < \beta_H(x)$ because $q_1(x) > q_2(x)$. For the second part, note that

$$\beta_L(x) = \frac{\alpha(1-\varphi(x))(1-q_1(x))}{\alpha(1-\varphi(x))(1-q_1(x)) + (1-\alpha)(1-q_2(x))} = \frac{a(x)}{a(x)+b(x)}, \tag{35}$$

where $a(\cdot)$ is decreasing in x and $b(x)$ is increasing.

(b) is a direct implication of utility maximization.

For (c), note that $P(\varepsilon_L|x) + P(\varepsilon_H|x) = \alpha(1-\varphi(x)) + (1-\alpha)$, so $[P(\varepsilon_L|x) + P(\varepsilon_H|x)]' < 0$ since $\varphi'(x) > 0$. ∎

Proof of Lemma 2. By the envelope theorem, the value functions in (15), (16), and

(19) are strictly decreasing in the penalty p.

For convexity, note that all three problems are of the form $f(p) = \max_{x \geq 0} n(x) - pm(x)$ and that, $x^*(p)$, the optimal effort x for a given value of p has the property that $\frac{d}{dp}x^*(p)$ has the opposite sign of $m'(x^*(p))$. Now, $f(p) = n(x^*(p)) - pm(x^*(p))$ where $x^*(p)$ satisfies $[n'(x^*(p)) - pm'(x^*(p))] \equiv 0$. Therefore, $f'(p) = [n'(x^*) - pm'(x^*)]\frac{d}{dp}x^*(p) - m(x^*(p))$ and $f''(p) = -m'(x^*)\frac{d}{dp}x^*(p)$. Since the two terms in this product have the opposite sign, we have $f''(p) \geq 0$.

We now show that the value functions in the two regimes have the "single-crossing from above in p" property. We treat the simpler case, regime $\mathcal{W}$, first. Since (T, T) is the agency's optimal policy at $p = 0$, $f_{\mathcal{W}}^{T,T}(0) > f_{\mathcal{W}}^{\neg T,T}(0)$. By the envelope theorem, under regime $\mathcal{W}$ with penalties $(p, 0)$, $\frac{d}{dp}f_{\mathcal{W}}^{T,T}(p) = -P(\varepsilon_L|x) < 0$, while $\frac{d}{dp}f_{\mathcal{W}}^{\neg T,T}(p) = 0$.

For the same reasons, under regime $\mathcal{B}(p, p)$, the crossings of $f_{\mathcal{B}}^{\neg T,\neg T}(p)$ happen from above. All that is left to consider is the crossing of $f_{\mathcal{B}}^{T,T}(p)$ and $f_{\mathcal{B}}^{\neg T,T}(p)$. In order to show that $f_{\mathcal{B}}^{T,T}(p)$ crosses $f_{\mathcal{B}}^{\neg T,T}(p)$ at most once from above as p increases, it is sufficient to show that for at any $p^\ddagger$ where $f_{\mathcal{B}}^{T,T}(p^\ddagger) = f_{\mathcal{B}}^{\neg T,T}(p^\ddagger)$, $f_{\mathcal{B}}^{T,T}(\cdot)$ is steeper than $f_{\mathcal{B}}^{\neg T,T}(\cdot)$. By the envelope theorem again, we need to show that

$$[P(\varepsilon_L|x^*_{T,T}(p^\ddagger)) + P(\varepsilon_H|x^*_{T,T}(p^\ddagger))] > P(\varepsilon_H|x^*_{\neg T,T}(p^\ddagger)), \tag{36}$$

where $x^*_{T,T} = x^*_{T,T}(p^\ddagger)$ and $x^*_{\neg T,T} = x^*_{\neg T,T}(p^\ddagger)$ are the agency's corresponding optimal effort levels under regime $\mathcal{B}(p^\ddagger, p^\ddagger)$ when following the torture policies (T, T) and $(\neg T, T)$. Now, at $p^\ddagger$, both torture policies $(\neg T, T)$ and (T, T) are optimal, which implies that $\beta_L(x^*_{\neg T,T}) \leq p^\ddagger/\theta\delta D \leq \beta_L(x^*_{T,T})$. Since $\beta'_L(x) < 0$, this implies that $x^*_{\neg T,T} \geq x^*_{T,T}$. Then,

$$\begin{aligned} &P(\varepsilon_L|x^*_{TT}) + P(\varepsilon_H|x^*_{TT}) - P(\varepsilon_H|x^*_{\neg TT}) \\ &= \alpha(1 - \varphi(x^*_{TT})) + (1 - \alpha) - \alpha(1 - \varphi(x^*_{\neg TT}))q_1(x^*_{\neg TT}) - (1 - \alpha)q_2(x^*_{\neg TT}) \end{aligned} \tag{37}$$

$$= (1-\alpha)\left[1-q_2(x^*_{\neg TT})\right] + \alpha\left\{1-\varphi(x^*_{TT}) - q_1(x^*_{\neg TT}) + \varphi(x^*_{\neg TT})q_1(x^*_{\neg TT})\right\}$$

$$= (>0) + \alpha\left\{1-\varphi(x^*_{\neg TT}) + \varphi(x^*_{\neg TT}) - \varphi(x^*_{TT}) - q_1(x^*_{\neg TT}) + \varphi(x^*_{\neg TT})q_1(x^*_{\neg TT})\right\}$$

$$= (>0) + \alpha\underbrace{\left\{1-\varphi(x^*_{\neg TT})\right\}\left\{1-q_1(x^*_{\neg TT})\right\}}_{>0} + \underbrace{\alpha\left\{\varphi(x^*_{\neg TT}) - \varphi(x^*_{TT})\right\}}_{\geq 0} > 0,$$

where the weak inequality comes from $x^*_{\neg TT} \geq x^*_{TT}$ and $\varphi'(x) > 0$. ∎

Proof of Proposition 2. From Lemma **2**, $f^{T,T}_{\mathcal{B}}(p)$ crosses $f^{\neg T,T}_{\mathcal{B}}(p)$ and $f^{\neg T,T}_{\mathcal{B}}(p)$ crosses $f^{\neg T,\neg T}_{\mathcal{B}}(p)$ once from above as p increases. Let $\overline{p_{\mathcal{B}}}$ and $\underline{p_{\mathcal{B}}}$ be the points at which $f^{T,T}_{\mathcal{B}}(\overline{p_{\mathcal{B}}}) = f^{\neg T,T}_{\mathcal{B}}(\overline{p_{\mathcal{B}}})$ and $f^{\neg T,T}_{\mathcal{B}}(\underline{p_{\mathcal{B}}}) = f^{\neg T,\neg T}_{\mathcal{B}}(\underline{p_{\mathcal{B}}})$, respectively. If $\underline{p_{\mathcal{B}}} < \overline{p_{\mathcal{B}}}$, under regime $\mathcal{B}(p^\circ, p^\circ)$, for $p^\circ \in [0, \underline{p_{\mathcal{B}}})$, $V_{\mathcal{B}}(p^\circ) = f^{T,T}_{\mathcal{B}}(p^\circ)$, for $p^\circ \in (\underline{p_{\mathcal{B}}}, \overline{p_{\mathcal{B}}})$, $V_{\mathcal{B}}(p^\circ) = f^{\neg T,T}_{\mathcal{B}}(p^\circ)$, and for $p^\circ > \overline{p_{\mathcal{B}}}$, $V_{\mathcal{B}}(p^\circ) = f^{\neg T,\neg T}_{\mathcal{B}}(p^\circ)$. The agency chooses $(\neg T, T)$ under regime $\mathcal{B}$ if $p^\circ \in (\underline{p_{\mathcal{B}}}, \overline{p_{\mathcal{B}}})$.

Similarly, from Lemma **2**, $f^{T,T}_{\mathcal{W}}(p)$ crosses $f^{\neg T,T}_{\mathcal{W}}(p)$ once from above as p increases. Let $p_{\mathcal{W}}$ be the point at which $f^{T,T}_{\mathcal{W}}(p_{\mathcal{W}}) = f^{\neg T,T}_{\mathcal{W}}(p_{\mathcal{W}})$. Under regime $\mathcal{W}(p^\circ, 0)$, for $p^\circ \in [0, p_{\mathcal{W}})$, $V_{\mathcal{W}}(p^\circ) = f^{T,T}_{\mathcal{W}}(p^\circ)$, and for $p^\circ > p_{\mathcal{W}}$, $V_{\mathcal{W}}(p^\circ) = f^{\neg T,T}_{\mathcal{W}}(p^\circ)$. The agency chooses $(\neg T, T)$ under regime $\mathcal{W}$ if $p^\circ > p_{\mathcal{W}}$.

Thus, if $p_{\mathcal{W}} < \overline{p_{\mathcal{B}}}$, then for $p^\circ \in (\max\{p_{\mathcal{W}}, \underline{p_{\mathcal{B}}}\}, \overline{p_{\mathcal{B}}})$, the agency chooses $(\neg T, T)$ under both regimes $\mathcal{B}$ and $\mathcal{W}$. ∎

Proof of Proposition 3. For $t \in \{0, 1\}$ and $x \geq 0$, define

$$h(x,t) = \left[\psi(x)\mathbb{D}(x) - c(x)\right] - p^\circ\left\{0 + t \cdot P(\varepsilon_H|x)\right\} \tag{38}$$

so that $\max_{x\geq 0} h(x,1)$ is the agency's optimal effort problem in regime $\mathcal{B}(p^\circ, p^\circ)$ if following the policy $(\neg T, T)$, and $\max_{x\geq 0} h(x,0)$ is the agency's optimal effort problem in regime $\mathcal{W}(p^\circ, 0)$ if following the policy $(\neg T, T)$. Simple increasing differences comparative statics (e.g., Corbae, Stinchcombe, and Zeeman, 2008, §2.8.b) show that $x^*(1) \geq x^*(0)$ if for all $x' > x$, $h(x',1) - h(x,1) > h(x',0) - h(x,0)$, and $x^*(1) \leq x^*(0)$ if for all $x' > x$, $h(x',1) -$

$h(x,1) < h(x',0) - h(x,0)$. Applied to (38), $h(x',1) - h(x,1) > h(x',0) - h(x,0)$ iff

$$-p^\circ \cdot \{P(\varepsilon_H|x') - P(\varepsilon_H|x)\} > 0. \tag{39}$$

Thus, $[P(\varepsilon_H|x)]' < 0$ for all x implies that $\mathcal{B}(p^\circ, p^\circ)$ has higher optimal effort than $\mathcal{W}(p^\circ, 0)$, part (a) of the Proposition, and $[P(\varepsilon_H|x)]' > 0$ for all x implies that $\mathcal{B}(p^\circ, p^\circ)$ has lower optimal effort than $\mathcal{W}(p^\circ, 0)$, part (b) of the Proposition. ∎

Proof of Proposition 4. In both (a) and (b), the switch from $\mathcal{B}(p^\circ, p^\circ)$ to $\mathcal{W}(p^\circ, 0)$ involves changing the policy from not torturing when evidence is low to torturing when evidence is low. By Lemma 1(a), $\beta'_L(x) < 0$. Therefore, by Lemma 1(b), the change in torture policy indicates a decrease in effort. The argument for the increasing frequency of torture of the innocent follows, as above, from $0 < q_2(x) < 1$. ∎

Proof of Proposition 5. From Lemma **2**, $f_{\mathcal{B}}^{T,T}(p)$ crosses $f_{\mathcal{B}}^{\neg T,T}(p)$ and $f_{\mathcal{B}}^{\neg T,T}(p)$ crosses $f_{\mathcal{B}}^{\neg T,\neg T}(p)$ once from above as p increases. Let $\overline{p_{\mathcal{B}}}$ and $\underline{p_{\mathcal{B}}}$ be the points at which $f_{\mathcal{B}}^{T,T}(\overline{p_{\mathcal{B}}}) = f_{\mathcal{B}}^{\neg T,T}(\overline{p_{\mathcal{B}}})$ and $f_{\mathcal{B}}^{\neg T,T}(\underline{p_{\mathcal{B}}}) = f_{\mathcal{B}}^{\neg T,\neg T}(\underline{p_{\mathcal{B}}})$, respectively. Under regime $\mathcal{B}(p^\circ, p^\circ)$, if $\underline{p_{\mathcal{B}}} < \overline{p_{\mathcal{B}}}$, then for $p^\circ \in (\underline{p_{\mathcal{B}}}, \overline{p_{\mathcal{B}}})$, $V_{\mathcal{B}}(p^\circ) = f_{\mathcal{B}}^{\neg T,T}(p^\circ)$, and thus, the agency chooses $(\neg T, T)$, and for $p > \overline{p_{\mathcal{B}}}$, $V_{\mathcal{B}}(p^\circ) = f_{\mathcal{B}}^{\neg T,\neg T}(p^\circ)$, and thus, the agency chooses $(\neg T, \neg T)$.

Similarly, from Lemma **2**, $f_{\mathcal{W}}^{T,T}(p)$ crosses $f_{\mathcal{W}}^{\neg T,T}(p)$ once from above as p increases. Let $p_{\mathcal{W}}$ be the point at which $f_{\mathcal{W}}^{T,T}(p_{\mathcal{W}}) = f_{\mathcal{W}}^{\neg T,T}(p_{\mathcal{W}})$. Under regime $\mathcal{W}(p^\circ, 0)$, for $p^\circ \in [0, p_{\mathcal{W}})$, $V_{\mathcal{W}}(p^\circ) = f_{\mathcal{W}}^{T,T}(p^\circ)$, and thus, the agency chooses (T, T).

Case (a): if $\overline{p_{\mathcal{B}}} < p_{\mathcal{W}}$, then the agency chooses $(\neg T, \neg T)$ under regime $\mathcal{B}$ and (T, T) under regime $\mathcal{W}$ if $p^\circ \in (\overline{p_{\mathcal{B}}}, p_{\mathcal{W}})$.

Case (b): if $\underline{p_{\mathcal{B}}} < p_{\mathcal{W}}$, then the agency chooses $(\neg T, T)$ under regime $\mathcal{B}$ and (T, T) under regime $\mathcal{W}$ if $p^\circ \in (\underline{p_{\mathcal{B}}}, \min\{\overline{p_{\mathcal{B}}}, p_{\mathcal{W}}\})$. ∎.

REFERENCES

Andreoni, James (1991). "Reasonable Doubt and the Optimal Magnitude of Fines: Should the Penalty Fit the Crime?" *RAND Journal of Economics* 22, 385-395.

Atkins, Raymond A. and Rubin, Paul H. (2003). "Effects of Criminal Procedure on Crime Rates: Mapping Out the Consequences of the Exclusionary Rule," *Journal of Law and Economics* 46, 157-180.

Bagaric, Mirko and Clarke, Julie (2006). *Torture: When the Unthinkable is Morally Permissible*, State University of New York Press: New York.

Berman, Eli and Laitin, David D. (2008). "Religion, Terrorism and Public Goods: Testing the Club Model," *Journal of Public Economics* 92, 1942-1967.

Chen, Kong-Ping, Tsai, Tsung-Sheng, and Leung, Angela (2009). "Judicial Torture as a Screening Device," Academia Sinica Working Paper.

Chen, Kong-Ping, Chou, Chien-Fu, and Tsai, Tsung-Sheng (2009). "Judicial Torture as a War of Attrition," Academia Sinica Working Paper.

Corbae, Dean, Stinchcombe, Maxwell B., and Zeeman, Juraj (2008). *An Introduction to Mathematical Analysis for Economic Theory and Econometrics*, Princeton University Press: New Jersey, Forthcoming.

Costanzo, Mark, Gerrity, Ellen, and Lykes, M. Brinton (2007). "Psychologists and the Use of Torture," *Analyses of Social Issues and Public Policy* 7, 7–20.

Dershowitz, Alan M. (2002). *Why Terrorism Works*, Yale University Press: New Haven.

Dershowitz, Alan M. (2003). "The Torture Warrant: A Response to Professor Strauss," *New York Law School Law Review* 48, 275-294.

Dreher, Axel, Gassebner, Martin, and Siemers, Lars-H. R. (2007). "Does Terror Threaten Human Rights? Evidence From Panel Data," CESifo Working Paper No. 1935.

Eggen, Dan (2007a). "Mukasey Losing Democrats' Backing; Nominee Unsure if Waterboarding Breaks Torture Law," *Washington Post* October 31.

Eggen, Dan (2007b). "Torture Stance Raises Doubts on Mukasey," *Washington Post* October 27.

Enders, Walter and Sandler, Todd (2004). "An Economic Perspective on Transnational Terrorism," *European Journal of Political Economy* 20, 301–316.

Enders, Walter and Sandler, Todd (2005). "Transnational Terrorism 1968-2000: Thresholds, Persistence, and Forecasts," *Southern Economic Journal* 71, 467-483.

Fay, Major General George R. (2004). "Fay Report: Investigation of Intelligence Activities At Abu Ghraib: Executive Summary," Available at Website http://fl1.findlaw.com/news.findlaw.com/hdocs/docs/dod/fay82504rpt.pdf.

Garoupa, Nuno, Klick, Jonathan, and Parisi, Francesco (2006). "A Law and Economics Perspective on Terrorism," *Public Choice* 128, 147-168.

Garoupa, Nuno M. (2007). "On the Optimal Choice of Enforcement Technology: An Efficiency Explanation of Privacy Rights," *Revue Économique* 56, 1353-1363.

Garrido, Eugenio, Masip, Jaume, and Herrero, Carmen (2004). "Police Officers' Credibility Judgments: Accuracy and Estimated Ability," *International Journal of Psychology* 39, 254–275.

Imseis, Ardi (2001). "'Moderate' Torture On Trial: The Israeli Supreme Court Judgement Concerning the Legality of the General Security Service Interrogation Methods," *International Journal of Human Rights* 5, 71-96.

Kaplow, Louis and Shavell, Steven (1996). "Property Rules versus Liability Rules: An Economic Analysis," *Harvard Law Review* 109, 713-790.

Kassin, Saul M, Goldstein, Christine C., and Savitsky, Kenneth (2003). "Behavioral Confirmation in the Interrogation Room: On the Dangers of Presuming Guilt," *Law and Human Behavior* 27, 187-203.

Kontorovich, Eugene (2004). "Liability Rules for Constitutional Rights: The Case of Mass Detentions," *Stanford Law Review* 56, 755-833.

Krueger, Alan B. and Maleckova, Jitka (2003). "Education, Poverty, and Terrorism: Is There a Causal Connection?" *Journal of Economic Perspectives* 17, 119-44.

Krueger, Alan B. (2007). *What Makes a Terrorist?* Princeton University Press: New Jersey.

Leshem, Shmuel (2009). "The Benefits of a Right to Silence for the Innocent," *RAND Journal of Economics*, Forthcoming.

Luban, David (2005). "Liberalism, Torture, and the Ticking Bomb," *Virginia Law Review* 91, 1425-1461.

Martinez, Jenny S. (2007). "The Military Commissions Act and 'Torture Light'," *Harvard International Law Journal* 48, 58-61.

Mialon, Hugo M. (2005). "An Economic Theory of the Fifth Amendment," *RAND Journal of Economics* 36, 834-849.

Mialon, Hugo M. and Mialon, Sue H. (2008). "The Effects of the Fourth Amendment: An Economic Analysis," *Journal of Law, Economics, and Organization* 24, 22-44.

Mialon, Hugo M. and Rubin, Paul H. (2008). "The Economics of the Bill of Rights," *American Law and Economics Review* 10, 1-60.

Mulligan, Casey B., Gil, Ricard, and Sala-i-Martin, Xavier (2004). "Do Democracies Have Different Public Policies than Nondemocracies?" *Journal of Economic Perspectives* 18, 51-74.

Persson, Mats and Siven, Claes-Henric (2007). "The Becker Paradox and Type I Versus Type II Errors in the Economics of Crime," *International Economic Review* 48, 211-233.

Posner, Richard (2002). "The Best Offense," *The New Republic* September 2, 28-31.

Rejali, Darius (2007). *Torture and Democracy*, New Jersey: Princeton University Press.

Rizzo, Mario J. and Whitman, Douglas Glen (2004). "The Camel's Nose is in the Tent: Rules, Theories, and Slippery Slopes," *UCLA Law Review* 51, 539-592.

Sandler, Todd and Siqueira, Kevin (2006). "Global Terrorism: Deterrence versus Preemption," *Canadian Journal of Economics* 50, 1370-1387.

Seidmann, Daniel J. and Stein, Alex (2000). "The Right to Silence Helps the Innocent: A Game-Theoretic Analysis of the Fifth Amendment Privilege," *Harvard Law Review* 114, 431-510.

Shavell, Steven (1991). "Specific versus General Enforcement of Law," *Journal of Political Economy* 99, 1088-1108.

Shavell, Steven (2007). "Optimal Discretion in the Application of Rules," *American Law and Economics Review* 9, 175-194.

Siqueira, Kevin and Sandler, Todd (2007). "Terrorist Backlash, Terrorism Mitigation, and Policy Delegation," *Journal of Public Economics* 91, 1800-1815.

Sobel, Joel (2000). "A Model of Declining Standards," *International Economic Review* 41, 295-303.

Sobel, Joel (2001). "On the Dynamics of Standards," *RAND Journal of Economics* 32, 606-623.

Stephenson, Matthew C. (2007). "Bureaucratic Decision Costs and Endogenous Agency Expertise," *Journal of Law, Economics, and Organization* 23, 469-498.

Strauss, Marcy (2003). "Torture," *New York Law School Law Review* 48, 201-274.

The Economist (2006), "Saying No to Torture," October 20, Available at Website http://www.economist.com/agenda/displaystory.cfm?story_id=8070066.

The Pew Research Foundation (2009), "No Change in Views on Torture, Warrantless Wiretaps," February 18, Available at Website http://people-press.org/reports/pdf/493.pdf.

Volokh, Eugene (2003). "The Mechanisms of the Slippery Slope," *Harvard Law Review* 116, 1026-1134.

Waldron, Jeremy (2005). "Torture and Positive Law: Jurisprudence for the White House," *Columbia Law Review* 105, 1681-1750.

Wantchekon, Leonard and Healy, Andrew (1999). "The 'Game' of Torture," *Journal of Conflict Resolution* 43, 596-609.

Wickelgren, Abraham L. (2010). "A Right to Silence for Civil Defendants?" *Journal of Law, Economics, and Organization* 26.