

Appendices for

Interrogational Torture: Or How Good Guys Get Bad Information with Ugly

Methods

John W. Schiemann

Fairleigh Dickinson University

\*Draft\*

6.26.2009

Please do not cite.

## Appendix A

This appendix contains the proofs and formal statements of the equilibria discussed in section three. I solve for pure strategy Perfect Bayesian equilibria. The game begins with two independent moves by nature. The first move selects the detainee's type,  $D_i$ , from the type space {Weak, Strong, Innocent}  $\equiv \{D_W, D_S, D_Z\}$  with the common prior probability distribution  $p_w, p_s,$  and  $p_z$ , where  $p_i$  is the probability the detainee is type  $i$ , and  $p_w+p_s+p_z=1$ . Nature's second move selects the state's type,  $S_i$ , from the type space {Pragmatic, Sadistic}  $\equiv \{S_P, S_S\}$  with the common prior probability distribution  $q_p$  and  $q_s$ , where  $q_i$  is the probability the state is type  $i$ , and  $q_p+q_s=1$ . Each  $D_i$  chooses a strategy from  $\{m,n\}$ . Strategies for  $D_i$  are given as  $(r_1,r_2,r_3)$  indicating that  $D_W$  chooses  $r_1$ ,  $D_S$  chooses  $r_2$ , and  $D_Z$  chooses  $r_3$ . Note that  $D_Z$  has move  $m$  under leading questioning only; under objective questioning  $D_Z$  has move  $n$  only. Each  $S_i$  chooses a strategy  $\{\tau,\tau\}$ , with  $(s_1,s_2)$  denoting that  $S_P$  chooses  $s_1$  when it observes  $m$  and chooses  $s_2$  when it observes  $n$  and likewise for  $S_S$  with  $(t_1,t_2)$ .

Both the weak and strong detainees pay costs  $i, i>0$  for  $m$  and receive a payoff of 0 for  $n$ . They also suffer costs  $k, k>0$  if they are tortured by the state and receive a payoff of 0 for no torture. The preference orderings for each are:  $D_W = \{0>-i>-k>-i-k\}$  and  $D_S = \{0>-k>-i>-i-k\}$ . The innocent detainee's payoff ordering is identical to that of the weak detainee, with  $\lambda$  taking the place of  $i$  for the cost of sending message  $m$ . Both state types pay a cost  $\sigma, \sigma>0$  if they fail to torture after receiving message  $n$  from a knowledgeable detainee and 0 for not torturing after message  $n$  from an innocent detainee.  $S_P$  bears a cost  $c, c>0$  for torturing any  $D_i$  and an additional cost  $\alpha, \alpha>0$  (with  $-c>-\sigma>-\alpha$ ), for "unnecessary" torture: of an innocent detainee who sends message  $n$  (i.e. tells the truth) or of any detainee who reveals full information with message  $m$ ). In contrast,  $S_S$  receives a benefit  $s, s>0$  to torture after any move by  $D_i$ .

Both states receive a payoff of  $V$  for detainee messages  $m$  that provide all the information they have to the state; for fractions less than full information, the states receive a payoff of  $V-\delta$ . Since the value of  $m$  is private information on the part of the (knowledgeable) detainee, the states' have only the common prior belief that  $m$  provides  $V$  with probability  $\theta$  and  $V-\delta$  with probability  $1-\theta$ , with  $\theta \in (0,1)$ . In the objective questioning variant of the model, the message  $m$  is perceived by  $S_P$  as  $m$  with probability  $\omega$

and is perceived as message  $n$  (a non-valuable message) with probability  $1-\omega$ ,  $\omega \in (0,1]$ . This uncertainty is the state's private information; the detainee assumes that the state recognizes any message  $m$  as valuable ( $\omega=1$ ) and plays accordingly.  $S_p$  assumes that the prior belief  $\omega$  is common knowledge and plays accordingly. Note both that the uncertainty captured by  $\omega$  occurs under objective questioning only – there is no uncertainty over the value of information under leading questioning – and that the state's belief about whether a message  $m$  is a valuable ( $\omega$ ) is independent of the state's belief about whether the detainee is hiding information ( $\theta$ ) (i.e.  $m$  is fractional rather than full).

Further, let  $\mu_{m,j} \in (0,1)$  denote  $S_i$ 's beliefs at her  $m$  information set, where  $\mu_{m,j}$  is the probability the detainee is type  $j$ ,  $j \in \{D_W, D_S, D_Z\}$ , after observing  $m$ . Similarly, let  $\mu_{n,k} \in (0,1)$  denote  $S_i$ 's beliefs at her  $n$  information set, where  $\mu_{n,k}$  is the probability the detainee is type  $k$ ,  $k \in \{D_W, D_S, D_Z\}$ , after observing  $n$ . Finally, I make the following knife-point assumptions to rule out indifference between strategy choices for  $D_i$  and  $S_i$ : If payoff-indifferent between sending messages  $m$  and  $n$ ,  $D_W$  and  $D_Z$  prefer to send  $m$ ; if payoff-indifferent between  $\tau$  and  $\sim\tau$ ,  $S_p$  prefers  $\sim\tau$ .

### A1. Objective Questioning

Under objective questioning the signal  $m$  is potentially noisy rather than clear, so  $S_p$ 's payoffs after  $m$  are weighted by  $\omega$ ,  $\omega \in (0,1]$  but  $\omega = 1$  for any  $D_i$  sending  $m$ . Since  $n$  dominates  $m$  for  $D_S$  and  $D_Z$  only has move  $n$  under objective questioning, there are only two pure strategies to consider,  $(m,n,n)$  and  $(n,n,n)$ .

#### Case O1: $\{m,n,n\}$

Suppose  $D_i$  plays the strategy  $(m,n,n)$ ; using Bayes's theorem,  $S_p$ 's beliefs at the  $m$  information set are  $\mu_m(D_W|m) = 1$ ,  $\mu_m(D_S|m) = 0$ ,  $\mu_m(D_Z|m) = 0$  and at the  $n$  information set are  $\mu_n(D_W|n) = 0$ ,  $\mu_n(D_S|n) = \frac{p_S}{p_S + p_Z}$ ,  $\mu_n(D_Z|n) = \frac{p_Z}{p_S + p_Z}$ . Given these beliefs, the expected utility of  $\tau$  at the  $m$  information set is  $\omega V - \omega \delta - \omega \theta \alpha + \omega \theta \delta - c$  and the expected utility of  $\sim\tau$  at the  $m$  information set is  $\omega V - \omega \delta + \omega \theta \delta + \omega \theta \sigma - \sigma$ .  $S_p$  therefore prefers to torture after  $m$  if

$$\omega < \frac{\sigma - c}{\theta\sigma + \theta\alpha} \equiv \hat{\omega} \quad (1)$$

or, solving for  $\theta$ ,

$$\theta < \frac{\sigma - c}{\omega\sigma + \omega\alpha} \equiv \hat{\theta}. \quad (2)$$

These are the *signal recognition* and *information hiding thresholds*, respectively. Recalling the detainee's assumption that any  $m$  is recognized with certainty ( $\omega=1$ ), it will be useful to define the detainee's belief about the state's information hiding threshold as

$$\theta < \frac{\sigma - c}{\sigma + \alpha} \equiv \theta^*. \quad (3)$$

$S_p$ 's expected utility for  $\tau$  at her  $n$  information set is  $-c - \frac{p_Z\alpha}{p_S + p_Z}$  and her expected utility for  $\sim\tau$  after

$n$  is  $-\frac{p_S\sigma}{p_S + p_Z}$ .  $S_p$  therefore plays  $\tau$  after  $n$  for

$$p_Z < \frac{\sigma - c}{c + \alpha} p_S \equiv p^* \quad (4)$$

This is the *innocent to strong detainee ratio* threshold. By simple inspection of (2) and (3), it is clear that  $\theta^* \leq \hat{\theta}$  for all  $\omega$ ,  $\omega \in (0, 1]$ . Equations (2), (3), and (4) thus define six subcases.

Subcase O1a(1):  $\theta^* < \theta < \hat{\theta}$  and  $p < p^*$ . For this combination of beliefs,  $S_p$  plays  $(\tau, \tau)$ .  $S_S$  always prefers torture to not torture. It remains to check whether  $(m, n, n)$  is  $D_i$ 's best response to these choices. The strategy  $n$  dominates  $m$  for  $D_S$  and under objective questioning  $n$  is  $D_Z$ 's only strategy so they will not deviate. Because  $D_W$  believes that  $\theta^* < \theta$ , he believes  $S_p$  plays  $\sim\tau$  rather than  $\tau$  after  $m$ . For  $D_W$ , the expected utility of  $m$  is  $qk - i - k$  and the expected utility of  $n$  is  $-k$ . Thus,  $D_W$  prefers  $m$  to  $n$  for

$$q \geq \frac{i}{k} \equiv \hat{q} \quad (5)$$

This is the *weak detainee's pragmatic state recognition threshold*. With no incentive to deviate to  $n$ , the strategy profile  $\{(m,n,n); (\tau, \tau), (\tau, \tau): q > \hat{q}, \theta^* < \theta < \hat{\theta}; (\mu_m, \mu_n)\}$  for  $\mu_{m,w}=1, \mu_n=p < p^*$ , constitutes a PBE. *This is the valuable information, surprise torture equilibrium*.

Subcase O1a(2):  $\theta < \theta^* \leq \hat{\theta}$  and  $p < p^*$ . For this combination of beliefs,  $S_p$  plays  $(\tau, \tau)$ .  $S_S$  always prefers torture to not torture. It remains to check whether  $(m,n,n)$  is  $D_i$ 's best response to these choices. The strategy  $n$  dominates  $m$  for  $D_S$  and under objective questioning  $n$  is  $D_Z$ 's only strategy so they will not deviate. Because  $\theta < \theta^*$ ,  $D_W$  would anticipate  $S_p$ 's response of  $\tau$  after  $m$ , providing  $D_W$  with an incentive to switch to  $n$  and this set of strategies and beliefs cannot constitute a PBE.

Subcase O1b:  $\theta > \hat{\theta} \geq \theta^*$  and  $p < p^*$ . For this combination of beliefs,  $S_p$  chooses  $(\sim\tau, \tau)$  and  $S_S$  chooses  $(\tau, \tau)$ . It remains to check whether  $(m,n,n)$  is  $D_i$ 's best response to these choices. From (5),  $D_W$  prefers  $m$  to  $n$  for  $q \geq \hat{q}$ . The strategy  $n$  dominates  $m$  for  $D_S$  and under objective questioning  $n$  is  $D_Z$ 's only strategy so they will not deviate. Thus, the strategy profile  $\{(m,n,n); (\sim\tau, \tau), (\tau, \tau): q \geq \hat{q}, \theta > \hat{\theta}; (\mu_m, \mu_n)\}$ , for  $\mu_{m,w}=1, \mu_n=p < p^*$  constitute a PBE. This is a *valuable information – no torture equilibrium*.

Subcase O1c(1):  $\theta^* < \theta < \hat{\theta}$  and  $p \geq p^*$ . For this combination of beliefs,  $S_p$  plays  $(\tau, \sim\tau)$ .  $S_S$  always prefers torture to not torture. It remains to check whether  $(m,n,n)$  is  $D_i$ 's best response to these choices. The strategy  $n$  dominates  $m$  for  $D_S$  and under objective questioning  $n$  is  $D_Z$ 's only strategy so they will not deviate. Because  $D_W$  believes that  $\theta^* < \theta$ , he believes  $S_p$  plays  $\sim\tau$  rather than  $\tau$  after  $m$ .  $D_W$  nevertheless has an incentive to deviate because  $S_p$  plays  $\sim\tau$  after  $n$ , making  $n$  preferable to  $m$  for any  $q$  and preventing this strategy profile and combination of beliefs from constituting a PBE.

Subcase O1c(2):  $\theta < \theta^* < \hat{\theta}$  and  $p \geq p^*$ . For this combination of beliefs,  $S_p$  plays  $(\tau, \sim\tau)$ .  $S_S$  always prefers torture to not torture. It remains to check whether  $(m,n,n)$  is  $D_i$ 's best response to these choices. The strategy  $n$  dominates  $m$  for  $D_S$  and under objective questioning  $n$  is  $D_Z$ 's only strategy so they will not deviate. Because  $\theta < \theta^*$ ,  $D_W$  would anticipate  $S_p$ 's response of  $\tau$  after  $m$ . Since  $S_p$  plays  $\sim\tau$

after  $n$ ,  $D_W$  has an incentive to deviate to  $n$  and so this strategy profile and belief combination cannot be part of a PBE.

Subcase O1d:  $\theta > \hat{\theta} \geq \theta^*$  and  $p \geq p^*$ . For this combination of beliefs,  $S_P$  plays  $(\sim\tau, \sim\tau)$ .  $S_S$  always prefers torture to not torture. Since  $S_P$  plays  $\sim\tau$  after  $n$ ,  $D_W$  has an incentive to deviate to  $n$  and so this strategy profile and belief combination cannot be part of a PBE.

**Case O2: {n,n,n}**

Suppose  $D_i$  plays the strategy  $(n,n,n)$ ; using Bayes's theorem,  $S_P$ 's beliefs at the  $n$  information set are  $p_w$ ,  $p_s$ , and  $p_z$ . Given these beliefs,  $S_P$ 's expected utility from  $\tau$  after  $n$  is  $-c - p_z\alpha$  and her expected utility from  $\sim\tau$  after  $n$  is  $-\sigma(p_w + p_s)$ . Thus  $S_P$  plays  $\tau$  after  $n$  for

$$p_z < \frac{\sigma - c}{\alpha + \sigma} \equiv \hat{p} \quad (6)$$

This is the *innocent detainee recognition threshold*, providing two cases.

Subcase O2a:  $p < \hat{p}$ . For this set of  $S_P$  beliefs,  $S_P$  plays  $\tau$ ;  $S_S$  chooses the dominant strategy  $\tau$ . It remains to check whether  $(n,n,n)$  is  $D_i$ 's best response to these choices. The strategy  $n$  dominates  $m$  for  $D_S$  and under objective questioning  $n$  is  $D_Z$ 's only strategy so they will not deviate. By equation (5),  $D_W$  prefers  $n$  to  $m$  for  $q < \hat{q}$  and so will not deviate. For  $q \geq \hat{q}$  and  $\theta^* < \theta$ , however,  $D_W$  expects  $S_P$  to play  $\sim\tau$  after  $m$  and so has an incentive to deviate to  $m$ . To prevent  $D_W$ 's deviation to  $m$ ,  $S_P$  would have to play  $\tau$  after  $m$ . Under objective questioning, only  $D_W$  can send message  $m$ , so  $\mu_m(D_W|m) = 1$ . This is identical to Case O1 above so the expected utility of  $\tau$  and  $\sim\tau$  are given by  $\omega V - \omega\delta - \omega\theta\alpha + \omega\theta\delta - c$  and  $\omega V - \omega\delta + \omega\theta\delta + \omega\theta\sigma - \sigma$ , respectively, and, from equation (2),  $S_P$  therefore prefers to torture after  $m$  if

$$\theta < \frac{\sigma - c}{\omega\sigma + \omega\alpha} \equiv \hat{\theta}. \quad (2)$$

Further, for this off-path move to prevent  $D_W$ 's deviation,  $D_W$  must believe that  $S_P$  will play  $\tau$  after  $m$ , that is,  $\theta < \theta^* \leq \hat{\theta}$ . Thus, the strategy profile  $\{(n,n,n); (\tau, \tau), (\tau, \tau): (q < \hat{q} \text{ or } q \geq \hat{q} \text{ and } \theta < \theta^* \leq \hat{\theta}); (\mu_m, \mu_n)\}$ , for  $\mu_{m,w} = 1$ , and  $\mu_n = p < \hat{p}$  is a PBE. This is the *no information – torture pooling equilibrium*.

Subcase O2b:  $p \geq \hat{p}$ . For this set of  $S_P$  beliefs,  $S_P$  plays  $\sim\tau$ ;  $S_S$  chooses the dominant strategy ( $\tau$ ,  $\tau$ ). It remains to check whether  $(n,n,n)$  is  $D_i$ 's best response to these choices. No  $D_i$  can do better and so the strategy profile  $\{(n,n,n); (s_1, \sim\tau), (\tau, \tau): (q \in (0,1); (\mu_m, \mu_n))\}$ , for  $\mu_m = 0$ , and  $\mu_n = p \geq \hat{p}$  is a PBE. This is the *no-information – no torture pooling equilibrium*.<sup>1</sup>

## A2. Leading Questioning

In this case the interrogator's approach is leading questioning, causing  $\omega$  to drop out of  $S_P$ 's payoffs and making strategy  $m$  now available to  $D_Z$ . Because  $n$  continues to dominate  $m$  for  $D_S$ , there are four pure strategies to consider:  $\{m,n,m\}$ ,  $\{m,n,n\}$ ,  $\{n,n,m\}$ , and  $\{n,n,n\}$ .

### Case L1: $\{m,n,m\}$

Suppose  $D_i$  plays the strategy  $(m,n,m)$ ; using Bayes's theorem,  $S_P$ 's beliefs at the  $m$  information

set are  $\mu_m(D_W|m) = \frac{P_W}{P_W + P_Z}$ ,  $\mu_m(D_S|m) = 0$ ,  $\mu_m(D_Z|m) = \frac{P_Z}{P_W + P_Z}$  and at the  $n$  information set are

$\mu_n(D_W|n) = 0$ ,  $\mu_n(D_S|n) = 1$ ,  $\mu_n(D_Z|n) = 0$ . Given these beliefs,  $S_P$ 's expected utility for  $\tau$  after  $m$  is

$$\frac{p_W V - p_W \delta - p_W c - p_W \theta \alpha + p_W \theta \delta - p_Z c - p_Z \alpha}{p_W + p_Z} \text{ while the expected utility for } \sim\tau$$

is  $\frac{p_W V - p_W \delta - p_W \sigma + p_W \theta \delta + p_W \theta \sigma}{p_W + p_Z}$ .  $S_P$  therefore prefers  $\tau$  after  $m$  for

$$\tilde{p} \equiv \frac{\sigma - c - \theta(\sigma + \alpha)}{c + \alpha} p_W \quad (7)$$

This is the *weak detainee—ambiguous information threshold*.  $S_P$ 's expected utilities after  $n$  are  $-c$  for  $\tau$  and  $-\sigma$  for  $\sim\tau$ , so  $S_P$  plays  $\tau$  after  $n$ . There are thus two cases based on  $\tilde{p}$ .

### Subcase L1a: $p < \tilde{p}$ .

For this set of beliefs,  $S_P$  plays  $(\tau, \tau)$ .  $S_S$  always prefers torture to not torture. It remains to check whether  $(m,n,m)$  is  $D_i$ 's best response to these choices. The strategy  $n$  dominates  $m$  for  $D_S$ . Both  $D_W$  and  $D_Z$ , however, can do better by switching to  $n$  for any  $q$  and this combination of beliefs and strategies cannot be part of a PBE.

Subcase L1b:  $p \geq \tilde{p}$ .

For this set of beliefs,  $S_p$  plays  $(\sim\tau, \tau)$ .  $S_s$  always prefers torture to not torture. It remains to check whether  $(m,n,m)$  is  $D_i$ 's best response to these choices. The strategy  $n$  dominates  $m$  for  $D_s$ . From (5) above we know that  $D_w$  prefers  $m$  to  $n$  for  $q \geq \hat{q}$ . For  $D_z$ , the expected utility of  $m$  is  $qk-\lambda-k$  and the expected utility of  $n$  is  $-k$ . Thus,  $D_z$  prefers  $m$  to  $n$  for

$$q > \frac{\lambda}{k} \equiv q^* \quad (8)$$

This is the *innocent detainee pragmatic state recognition threshold*. Thus, the strategy profile  $\{(m,n,m); (\sim\tau, \tau), (\tau, \tau): q \geq \hat{q}, \text{ and } q \geq q^*; (\mu_m, \mu_n)\}$ , for  $\mu_m = p \geq \tilde{p}$  and  $\mu_{n,s}=1$ , is a PBE. This is the *ambiguous information equilibrium*.

### Case L2: $\{n,n,m\}$

Suppose  $D_i$  plays the strategy  $(n,n,m)$ ; using Bayes's theorem,  $S_p$ 's beliefs at the  $m$  information set are  $\mu_m(D_w|m) = 0, \mu_m(D_s|m) = 0, \mu_m(D_z|m) = 1$  and at the  $n$  information set are  $\mu_n(D_w|n) = \frac{p_w}{p_w + p_s}, \mu_n(D_s|n) = \frac{p_s}{p_w + p_s}, \mu_n(D_z|n) = 0$ . Given these beliefs,  $S_p$ 's expected utility for  $\tau$  after  $m$  is  $V-c-\alpha$  and his expected utility for  $\sim\tau$  is  $V$ .  $S_p$ 's expected utility for  $\tau$  after  $n$  is  $-c$  and his expected utility for  $\sim\tau$  is  $-\sigma$ , so  $S_p$  chooses  $(\sim\tau, \tau)$ .  $S_s$  chooses  $(\tau, \tau)$ . It remains to check whether  $(n,n,m)$  is  $D_i$ 's best response to these choices. From equation (5)  $D_w$  prefers  $n$  to  $m$  for  $q < \hat{q}$  and, from case O1a(2) above, for  $q \geq \hat{q}$  and  $\theta < \theta^* \leq \hat{\theta}$ . The strategy  $n$  dominates  $m$  for  $D_s$ . From case L1b immediately above,  $D_z$  prefers  $m$  to  $n$  for  $q \geq q^*$ . Thus, the strategy profile  $\{(n,n,m); (\sim\tau, \tau), (\tau, \tau): q < \hat{q}, q \geq \hat{q} \text{ and } \theta < \theta^* \leq \hat{\theta}, \text{ and } q \geq q^*; (\mu_m, \mu_n)\}$ , for  $\mu_{m,z}=1$  and  $\mu_{n,w} = \frac{p_w}{p_w + p_s}, \mu_{n,s} = \frac{p_s}{p_w + p_s}$ , constitute a PBE. This is a *false-confession equilibrium*.

### Case L3: $\{m,n,n\}$

This set of strategies on the part of  $D_i$  is identical to Case O1 above in which  $D_Z$  had move  $n$  only. Recalling that  $\omega$  drops from  $S_P$ 's payoffs under leading questioning, it remains only to check whether  $D_Z$  has an incentive to switch his strategy from  $n$  to  $m$  in any of the subcases in O1. There is only one subcase to consider because O1a(1) collapses to O1a(2) given that  $\omega=1 \Rightarrow \theta^* = \hat{\theta}$  and there was no equilibrium in O1a(2) and because there was no equilibrium in the last three subcases: O1c(1),(2), and O1d.

Subcase L3a:  $\theta > \hat{\theta} \geq \theta^*$  and  $p < p^*$ . For this combination of beliefs,  $S_P$  chooses  $(\sim\tau, \tau)$  and  $S_S$  chooses  $(\tau, \tau)$ . It remains to check whether  $(m,n,n)$  is  $D_i$ 's best response to these choices. From equation (8),  $D_Z$  prefers  $n$  to  $m$  for  $q < q^*$ . Thus, the strategy profile  $\{(m,n,n); (\sim\tau, \tau), (\tau, \tau): q \geq \hat{q}, q < q^*, \theta > \hat{\theta} \geq \theta^*; (\mu_m, \mu_n)\}$ , for  $\mu_{m/w}=1$  and  $\mu_n = p < p^*$  constitute a PBE. This is a *valuable information – no torture equilibrium*.

#### **Case 4: {n,n,n}**

Once again, this strategy profile is identical to its counterpart under objective questioning in Case O2 above, but, given that  $D_Z$  now has move  $m$  in addition to move  $n$ , it is necessary to check whether  $D_Z$  would deviate in each of the two subcases of O2 defined by equation (6).

Subcase L4a:  $p < \hat{p}$ . For this set of  $S_P$  beliefs,  $S_P$  plays  $\tau$ ;  $S_S$  chooses the dominant strategy  $\tau$ . It remains to check whether  $n$  is the best response for both  $D_W$  and  $D_Z$  under leading questioning. By equation (5),  $D_W$  prefers  $n$  to  $m$  for  $q < \hat{q}$  and so will not deviate; the same is true for  $D_Z$  for  $q < q^*$ . For  $q \geq \hat{q}$ ,  $q \geq q^*$ , and  $\theta < \theta^* \leq \hat{\theta}$   $D_W$  and  $D_Z$  expect  $S_P$  to play  $\tau$  after  $m$  and so will not deviate to  $m$ . For  $q \geq \hat{q}$ ,  $q \geq q^*$ , and  $\theta^* < \theta < \hat{\theta}$ , however,  $D_W$  and  $D_Z$  expect  $S_P$  to play  $\sim\tau$  after  $m$  and so have an incentive to deviate to  $m$ . To prevent deviation to  $m$  by  $D_W$  and  $D_Z$ ,  $S_P$  would have to play  $\tau$  after  $m$ . Since under leading questioning, both  $D_W$  and  $D_Z$  can send message  $m$  but  $D_S$  never does so, let  $\mu_{m,w}$  be  $S_P$ 's belief that the detainee is  $D_W$ , and  $1-\mu_{m,z}$  be  $S_P$ 's belief that the detainee is  $D_Z$ , upon observing  $m$ .  $S_P$  would therefore play  $\tau$  after  $m$  for

$$\mu_{m,w} > \frac{c + \alpha}{\alpha + \theta\alpha + \sigma + \theta\sigma} \equiv \mu_{m,w}^* . \quad (9)$$

Thus, the strategy profile  $\{(n,n,n); (\tau, \tau), (\tau, \tau): (q < \hat{q} \text{ and } q < q^* \text{ or } q \geq \hat{q} \text{ and } q \geq q^* \text{ and } \theta < \theta^* \leq \hat{\theta} \text{ or } \theta^* < \theta < \hat{\theta}); (\mu_m, \mu_n)\}$ , for  $\mu_m > \mu_m^*$  and  $\mu_n = p < \hat{p}$  is a PBE. This is the *no information – torture pooling equilibrium*.

Subcase L2b:  $p \geq \hat{p}$ . For this set of  $S_P$  beliefs,  $S_P$  plays  $(s_1, \sim\tau)$ ;  $S_S$  chooses the dominant strategy  $(\tau, \tau)$ . It remains to check whether  $(n,n,n)$  is  $D_i$ 's best response to these choices. No  $D_i$  can do better and so the strategy profile  $\{(n,n,n); (s_1, \sim\tau), (\tau, \tau): q \in (0,1); (\mu_m, \mu_n)\}$ , for  $\mu_n = p \geq \hat{p}$ , is a PBE. This is the *no information – no torture pooling equilibrium*.<sup>2</sup> Q.E.D.

## Appendix B

This appendix supports the comparative statics results in section three of the text by establishing the signs of the derivatives of the thresholds in appendix A with respect to their component parameters.

1. From equation (5) in appendix A,  $\hat{q} = \frac{i}{k}$ .

a.  $\frac{\partial \hat{q}}{\partial i} = \frac{1}{k} > 0$ .

b.  $\frac{\partial \hat{q}}{\partial k} = -\frac{1}{k^2} < 0$ .

2. From equation (2),  $\hat{\theta} = \frac{\sigma - c}{\omega\sigma + \omega\alpha}$ .

a.  $\frac{\partial \hat{\theta}}{\partial \sigma} = \frac{1}{\omega\sigma + \omega\alpha} - \frac{(\sigma - c)\omega}{(\omega\sigma + \omega\alpha)^2} > 0$ . Proof.  $\frac{\partial \hat{\theta}}{\partial \sigma} > 0 \Leftrightarrow \frac{1}{\omega\sigma + \omega\alpha} > \frac{\omega(\sigma - c)}{(\omega\sigma + \omega\alpha)^2} \Leftrightarrow$

$1 > \frac{\omega(\sigma - c)}{\omega\sigma + \omega\alpha} \Leftrightarrow 1 > \frac{\sigma - c}{\sigma + \alpha} \Leftrightarrow \sigma + \alpha > \sigma - c \Leftrightarrow \alpha > -c$ , which is true for  $0 < c < \sigma < \alpha$ .

b.  $\frac{\partial \hat{\theta}}{\partial c} = -\frac{1}{\omega\sigma + \omega\alpha} < 0$ .

$$c. \frac{\partial \hat{\theta}}{\partial \alpha} = -\frac{\omega(\sigma - c)}{(\omega\sigma + \omega\alpha)^2} < 0.$$

$$d. \frac{\partial \hat{\theta}}{\partial \omega} = -\frac{(\sigma - c)(\sigma + \alpha)}{(\omega\sigma + \omega\alpha)^2} < 0.$$

$$3. \text{ From equation (4), } p^* = \frac{\sigma - c}{c + \alpha} p_s$$

$$a. \frac{\partial p^*}{\partial \sigma} = \frac{p_s}{\alpha + c} > 0$$

$$b. \frac{\partial p^*}{\partial c} = -\frac{p_s}{\alpha + c} - \frac{p_s(\sigma - c)}{(\alpha + c)^2} < 0. \text{ Proof. } \frac{\partial p^*}{\partial c} < 0 \Leftrightarrow \frac{p_s}{\alpha + c} + \frac{p_s(\sigma - c)}{(\alpha + c)^2} > 0. \text{ Since both } \frac{p_s}{\alpha + c} \text{ and}$$

$\frac{p_s(\sigma - c)}{(\alpha + c)^2}$  are positive for  $0 < c < \sigma < \alpha$ , this must always hold.

$$c. \frac{\partial p^*}{\partial \alpha} = -\frac{p_s(\sigma - c)}{(\alpha + c)^2} < 0.$$

$$d. \frac{\partial p^*}{\partial p_s} = \frac{\sigma - c}{\alpha + c} > 0.$$

$$4. \text{ From equation (6), } \tilde{p} = \frac{\hat{p} - \theta}{r} p_w$$

$$a. \frac{\partial \tilde{p}}{\partial \theta} = -\frac{p_w}{r} < 0.$$

$$b. \frac{\partial \tilde{p}}{\partial r} = -\frac{(\hat{p} - \theta)p_w}{r^2} < 0. \text{ This holds for } \hat{p} > \theta.$$

$$c. \frac{\partial \tilde{p}}{\partial p_w} = \frac{\hat{p} - \theta}{r} > 0. \text{ This holds for } \hat{p} > \theta.$$

$$d. \frac{\partial \tilde{p}}{\partial \hat{p}} = \frac{p_w}{r} > 0.$$

## Appendix C

This appendix contains the proofs and formal statements of propositions one and two. For convenience, the appendix begins with a restatement of the relevant thresholds and six observations.

### Background, Definition, and Observations

From the derivations of the equilibria in appendix A, the relevant thresholds are:

1) From equation (4),  $p^* \equiv \frac{\sigma - c}{\alpha + c} p_S$ ,

2) From equation (6),  $\hat{p} \equiv \frac{\sigma - c}{\sigma + \alpha}$ , and

3) From equation (7),  $\tilde{p} \equiv \frac{\sigma - c - \theta(\sigma + \alpha)}{c + \alpha} p_W$ ,

where  $0 < c < \sigma < \alpha$ ,  $\theta \in (0,1)$ , and  $p_W, p_S, p_Z \in (0,1)$  with  $p_W + p_S + p_Z = 1$  and  $p_W > p_S \geq p_Z$  or  $p_W > p_Z \geq p_S$ .

Observation 1:  $\hat{p} < \frac{1}{2}$ . Proof.  $\hat{p} = \frac{\sigma - c}{\sigma + \alpha}$ . Given  $\alpha > \sigma > c$ ,  $\hat{p} = \frac{\sigma - c}{\sigma + \alpha} < \frac{\sigma}{2\sigma} = \frac{1}{2}$ .

Observation 2: Let  $r = \frac{\alpha + c}{\alpha + \sigma}$ . Then  $\frac{1}{2} < r < 1$ . Proof. It follows immediately that  $r < 1$  since  $c < \sigma$ .

Since  $\alpha > \sigma$ , let  $\alpha = \sigma + \varepsilon, \varepsilon > 0$ . Then  $r \equiv \frac{\alpha + c}{\alpha + \sigma} = \frac{\sigma + \varepsilon + c}{\sigma + \varepsilon + \sigma} \Leftrightarrow \frac{\sigma + \varepsilon + c}{2\sigma + \varepsilon} > \frac{\sigma}{2\sigma} = \frac{1}{2}$ .

Observation 3:  $\tilde{p} = \frac{\sigma - c - \theta(\sigma + \alpha)}{\alpha + c} p_W \Leftrightarrow \tilde{p} = \frac{\sigma - c}{\alpha + c} - \frac{\theta(\sigma + \alpha)}{\alpha + c} p_W \Leftrightarrow \tilde{p} = \frac{\hat{p} - \theta}{r} p_W$ .

Observation 4:  $\tilde{p} > 0$  iff  $0 < \theta < \hat{p}$ . This follows immediately from observation 3.

Observation 5: If  $p_W > p_S \geq p_Z$  or if  $p_W > p_Z \geq p_S$ , then  $p_S < \frac{1}{2}$ . Proof. Suppose  $p_S \geq \frac{1}{2}$ . Since

$p_W > p_S$ , this implies  $p_W > \frac{1}{2}$ . But this means  $p_W + p_S > 1$  which contradicts  $p_W + p_S + p_Z = 1$ .

Observation 6:  $\hat{p} \leq \hat{\theta}$ . Proof.  $\hat{p} \leq \hat{\theta} \Leftrightarrow \frac{\sigma - c}{\sigma + \alpha} \leq \frac{\sigma - c}{\omega\sigma + \omega\alpha}$ . It follows immediately that  $\hat{p} \leq \hat{\theta}$  for all

$\omega \in (0,1]$ .

**Proposition 1: If  $p_w > p_s \geq p_z$  or  $p_w > p_s \leq p_z$ , then  $p^* < \hat{p} < \frac{1}{2}$ .**

Lemma 1:  $p^* < \hat{p}$  iff  $0 < p_s < r$ .

$$\text{Proof. } p^* < \hat{p} \Leftrightarrow \frac{\sigma - c}{\alpha + c} p_s < \frac{\sigma - c}{\sigma + \alpha} \Leftrightarrow p_s < \frac{\frac{\sigma - c}{\sigma + \alpha}}{\frac{\sigma - c}{\alpha + c}} \Leftrightarrow p_s < \frac{\sigma - c}{\sigma + \alpha} \cdot \frac{\alpha + c}{\sigma - c} \Leftrightarrow$$

$$p_s < \frac{\alpha + c}{\sigma + \alpha} \equiv r.$$

Observation 1,  $\hat{p} < \frac{1}{2}$ , and Observation 2,  $r > \frac{1}{2}$ , together imply  $0 < \hat{p} < \frac{1}{2} < r < 1$ . From

Lemma 1,  $p^* > \hat{p}$  iff  $r < p_s < 1$ . This implies  $0 < \hat{p} < \frac{1}{2} < r < p_s < 1$  or  $p_s > \frac{1}{2}$ . Since, by

observation 5,  $p_s < \frac{1}{2}$ , then  $0 < p_s < r$  and  $p^* < \hat{p} < \frac{1}{2}$ . Q.E.D.

**Proposition 2: If the TJR exists under objective questioning, it is nowhere unique and shares its entire space with the ambiguous information, no torture equilibrium.**

Formally, this proposition amounts to identifying the necessary and sufficient conditions for the *valuable information, no torture equilibrium* under objective conditioning to exist and then showing that these conditions: 1) contradict the conditions for this equilibrium to be unique and 2) that they are also sufficient for the existence of the *ambiguous information, no torture equilibrium*. Or:

*The condition  $\theta > \hat{\theta}$  supporting the valuable information, no torture equilibrium under objective questioning contradicts the necessary and sufficient conditions for this equilibrium to be unique:*

either  $p_S < \left(1 - \frac{\theta}{\hat{p}}\right)p_W < r$  or  $p_S < \left(1 - \frac{\theta}{\tilde{p}}\right)p_W$  and  $p_W < r$  for  $p^* < \tilde{p} < \hat{p}$  or

$p_S < r < \left(1 - \frac{\theta}{\hat{p}}\right)p_W$  for  $p^* < \hat{p} < \tilde{p}$ . Further, the conditions supporting the valuable

information, no torture equilibrium *under objective questioning* also support the ambiguous information, no torture equilibrium.

Discussion in the main text showed that the valuable information, no torture equilibrium is unique only for  $p^* < \tilde{p} < \hat{p}$  and  $p^* < \hat{p} < \tilde{p}$ . The proof proceeds from this point in four steps. Part I establishes the necessary and sufficient conditions for  $p^* < \tilde{p} < \hat{p}$ . Part II establishes the necessary and sufficient conditions for  $p^* < \hat{p} < \tilde{p}$ . Part III demonstrates that that these conditions contradict the conditions supporting the valuable information, no torture equilibrium. Part IV establishes that the conditions supporting the valuable information, no torture equilibrium under objective questioning also support the *ambiguous information, no torture equilibrium*.

**Part I:** If  $p^* < \tilde{p} < \hat{p}$ , then the condition  $p_S < \left(1 - \frac{\theta}{\hat{p}}\right)p_W < r$  is necessary and sufficient to make the

valuable information, no torture equilibrium under objective questioning unique.

Lemma 2:  $p^* < \tilde{p}$  iff  $p_S < \left(1 - \frac{\theta}{\hat{p}}\right)p_W$ .

$$\text{Proof. } p^* < \tilde{p} \Leftrightarrow \frac{\sigma - c}{c + \alpha} p_S < \frac{\sigma - c - \theta(\sigma + \alpha)}{c + \alpha} p_W \Leftrightarrow (\sigma - c)p_S < \sigma - c - \theta(\sigma + \alpha)p_W \Leftrightarrow$$

$$p_S < \frac{\sigma - c - \theta(\sigma + \alpha)}{\sigma - c} p_W \Leftrightarrow p_S < \left(\frac{\sigma - c}{\sigma - c} - \frac{\theta(\sigma + \alpha)}{\sigma - c}\right)p_W \Leftrightarrow$$

$$p_S < \left(1 - \frac{\theta}{\frac{\sigma - c}{\sigma + \alpha}}\right)p_W \Leftrightarrow p_S < \left(1 - \frac{\theta}{\hat{p}}\right)p_W.$$

Lemma 3:  $\tilde{p} < \hat{p}$  iff  $r > \left(1 - \frac{\theta}{\hat{p}}\right)p_w$  or  $p_w < r$ .

Proof.

$$\hat{p} > \tilde{p} \Leftrightarrow \frac{\sigma - c}{\sigma + \alpha} > \frac{\sigma - c - \theta(\sigma + \alpha)}{\alpha + c} p_w \Leftrightarrow \frac{\sigma - c}{\sigma + \alpha} \cdot \frac{1}{p_w} > \frac{\sigma - c}{\alpha + c} - \frac{(\sigma + \alpha)}{\alpha + c} \theta \Leftrightarrow$$

$$\frac{\sigma + \alpha}{\alpha + c} \theta > \frac{\sigma - c}{\alpha + c} - \frac{\sigma - c}{\sigma + \alpha} \cdot \frac{1}{p_w} \Leftrightarrow \theta > \frac{\frac{\sigma - c}{\alpha + c}}{\frac{\alpha + \sigma}{\alpha + c}} - \frac{\frac{\sigma - c}{\alpha + \sigma} \cdot \frac{1}{p_w}}{\frac{\alpha + \sigma}{\alpha + c}} \Leftrightarrow$$

$$\theta > \frac{\sigma - c}{\alpha + c} \frac{\alpha + c}{\alpha + \sigma} - \frac{\sigma - c}{\alpha + \sigma} \frac{1}{p_w} \frac{\alpha + c}{\alpha + \sigma} \Leftrightarrow \theta > \frac{\sigma - c}{\alpha + \sigma} - \frac{\sigma - c}{\alpha + \sigma} \frac{1}{p_w} \frac{\alpha + c}{\alpha + \sigma} \Leftrightarrow$$

$$\frac{\theta}{\frac{\sigma - c}{\alpha + \sigma}} > \frac{\frac{\sigma - c}{\alpha + \sigma}}{\frac{\sigma - c}{\alpha + \sigma}} - \frac{\frac{\sigma - c}{\alpha + \sigma} \cdot \frac{1}{p_w} \frac{\alpha + c}{\alpha + \sigma}}{\frac{\sigma - c}{\alpha + \sigma}} \Leftrightarrow \frac{\theta}{\frac{\sigma - c}{\alpha + \sigma}} > 1 - \frac{1}{p_w} \frac{\alpha + c}{\alpha + \sigma} \Leftrightarrow$$

$$\left(\frac{\sigma - c}{\alpha + \sigma}\right) \frac{\theta}{\frac{\sigma - c}{\alpha + \sigma}} > \frac{\sigma - c}{\alpha + \sigma} \left(1 - \frac{1}{p_w} \frac{\alpha + c}{\alpha + \sigma}\right) \Leftrightarrow \theta > \frac{\sigma - c}{\alpha + \sigma} \left(1 - \frac{\alpha + c}{\alpha + \sigma} \frac{1}{p_w}\right) \Leftrightarrow \theta > \hat{p} \left(1 - \frac{r}{p_w}\right) \Leftrightarrow$$

$$\frac{\theta}{\hat{p}} > 1 - \frac{r}{p_w} \Leftrightarrow \frac{r}{p_w} > 1 - \frac{\theta}{\hat{p}} \Leftrightarrow r > \left(1 - \frac{\theta}{\hat{p}}\right)p_w.$$

Since, by Lemma 2,  $p^* < \tilde{p}$  iff  $p_s < \left(1 - \frac{\theta}{\hat{p}}\right)p_w$  and by Lemma 3,  $\tilde{p} < \hat{p}$  iff  $r > \left(1 - \frac{\theta}{\hat{p}}\right)p_w$  or

$p_w < r$  we obtain  $p^* < \tilde{p} < \hat{p}$  iff  $p_s < \left(1 - \frac{\theta}{\hat{p}}\right)p_w < r$  or  $p_s < \left(1 - \frac{\theta}{\hat{p}}\right)p_w$  and  $p_w < r$ . These are

the necessary and sufficient conditions to make the valuable information, no torture equilibrium under objective questioning unique for  $p^* < \tilde{p} < \hat{p}$ .

**Part II:** If  $p^* < \hat{p} < \tilde{p}$ , then the condition  $p_s < r < \left(1 - \frac{\theta}{\hat{p}}\right)p_w$  is necessary and sufficient to make the

valuable information, no torture equilibrium under objective questioning unique.

Since, from Lemma 1,  $p^* < \hat{p}$  iff  $0 < p_s < r$  and from Lemma 3,  $\tilde{p} > \hat{p}$  iff  $r < \left(1 - \frac{\theta}{\hat{p}}\right)p_w$ , we

obtain  $p^* < \hat{p} < \tilde{p}$  iff  $p_s < r < \left(1 - \frac{\theta}{\hat{p}}\right)p_w$ . Thus  $p_s < r < \left(1 - \frac{\theta}{\hat{p}}\right)p_w$  is necessary and sufficient to

make the valuable information, no torture equilibrium under objective questioning unique

for  $p^* < \hat{p} < \tilde{p}$ .

**Part III:**  $p_s < \left(1 - \frac{\theta}{\hat{p}}\right)p_w < r$  or  $p_s < \left(1 - \frac{\theta}{\hat{p}}\right)p_w$  and  $p_w < r$  and  $p_s < r < \left(1 - \frac{\theta}{\hat{p}}\right)p_w$  contradict

the conditions supporting the valuable information, no torture equilibrium.

*Proof.* The requirements for the orderings making the TJR unique are:  $p_s < \left(1 - \frac{\theta}{\hat{p}}\right)p_w < r$  or

$p_s < \left(1 - \frac{\theta}{\hat{p}}\right)p_w$  and  $p_w < r$  for  $p^* < \tilde{p} < \hat{p}$  and  $p_s < r < \left(1 - \frac{\theta}{\hat{p}}\right)p_w$ , for  $p^* < \hat{p} < \tilde{p}$ . Both

orderings thus require  $\theta < \hat{p} \equiv \frac{\sigma - c}{\sigma + \alpha}$ .

The *valuable information, no torture equilibrium* requires  $\theta > \hat{\theta} \equiv \frac{\sigma - c}{\omega\sigma + \omega\alpha}$ . Since, however, by

observation 6,  $\hat{p} \leq \hat{\theta}$ , this equilibrium requires  $\theta \geq \hat{p}$ , which contradicts the requirement  $\theta < \hat{p}$  for it to be unique.

**Part IV:** The conditions supporting the *valuable information, no torture equilibrium* under objective questioning also support the *ambiguous information, no torture equilibrium*.

From subcase O1b in appendix A, the relevant conditions on beliefs supporting the *valuable information, no torture equilibrium* under objective questioning are  $q \geq \hat{q}$ ,  $\theta > \hat{\theta}$ , and  $p < p^*$ . The conditions supporting the *ambiguous information, no torture equilibrium* are  $q \geq \hat{q}$ ,  $q \geq q^*$ , and

$p \geq \tilde{p}$ . The weak detainee thresholds  $\hat{q}$  are the same and the innocent detainee threshold  $q^*$  is irrelevant to the conditions supporting the *valuable information, no torture equilibrium* under objective questioning.

Consider  $p$ . The condition on  $p$  for the *valuable information, no torture equilibrium* under objective questioning is  $0 < p < p^*$ . The condition on  $p$  for the *ambiguous information, no torture equilibrium* is  $p \geq \tilde{p}$ . By observation 4,  $\tilde{p} > 0$  iff  $0 < \theta < \hat{p}$ . The *valuable information, no torture equilibrium* additionally requires  $\hat{\theta} < \theta$ . Since, by observation 6,  $\hat{p} = \hat{\theta}$  for  $\omega = 1$  (as is the case under leading questioning), the *valuable information, no torture equilibrium* thus requires  $\hat{p} = \hat{\theta} < \theta$  or  $\theta > \hat{p}$ . But this violates the condition that  $\tilde{p} > 0$  iff  $0 < \theta < \hat{p}$  in observation 4, so all  $p$  in the  $0 < p < p^*$  condition for the *valuable information, no torture equilibrium* under objective questioning are contained in the  $p \geq \tilde{p}$  condition for the *ambiguous information, no torture equilibrium*. Q.E.D.

## Notes

---

<sup>1</sup> Technically these conditions define two behaviorally equivalent equilibria, one in which  $s_1 = \sim\tau$  and one in which  $s_1 = \tau$ .

<sup>2</sup> Technically these conditions define two behaviorally equivalent equilibria, one in which  $s_1 = \sim\tau$  and one in which  $s_1 = \tau$ .