Conflict and Cooperation: Terrorism and Counterterrorism

Sandeep Baliga and Tomas Sjöström
Northwestern and Rutgers

Terrorism and Club Goods: Iannaccone and Berman

- Hamas is Islamic and has a terrorist wing. Tamil Tigers are not religious.
- *Both* provide public goods to members....

- Group members derive utility from (secular) consumption, S, and from time spent in religious activities, R, such as prayer and community service and from the level of a local public good A:

$$U(S_i, R_i, A) \; i = 1, 2, .., N$$

- Good A is *nonrival* and *excludable*, a *club* good.
- Members get A from either a government, G, or the "club," C, which uses hours of religious activity R_i as an input.
- Members maximize utility subject to time and budget constraints. A fixed allocation of time, T, is split between the religious activity, R_i, and work hours, H_i.
- Budget constraint: $pS_i = wH_i = w\left(T - R_i\right)$, p is price of secular good and w is wage.

- Let R^*, S^* maximize

$$U(\frac{w}{p}(T - R^*), R^*, NR^*)$$

so

$$U_2(T - R^*, R^*, NR^*) - \frac{w}{p}U_1(T - R^*, R^*, NR^*) + NU_3(T - R^*, R^*, NR^*) = 0$$

where we assume $U_i > 0$ and $U_{ii} < 0$ $i = 1, 2, 3$. Let U^* be the equilibrium utility.

- But an individual ignores the positive externality and maximizes

$$U(T - R_i, R_i, \Sigma_j R_j + R_i)$$

and undersupplies religious effort. Let $\tilde{R}$ be the equilibrium religious effort and $\tilde{U}$ the equilibrium utility.

Prohibition

- Sabbath restrictions and dress code: Suppose the club/sect forbids secular activity greater than $S^* = T - R^*$. This implements first-best. As $U^* > \tilde{U}$, the club will attract members.

Sacrifice

- Attending a seminary.
- Suppose there are high-wage H and low-wage types L. High wage types are bigger free-riders.
- We can have a segregated equilibrium: (1) there is one club L that has just low wage types who provide religious effort at level $\tilde{R}_L$ but require a costly sacrifice of time κ to join; (2) there is a high wage type club where members provide religious effort at level $\tilde{R}_H$ where $\tilde{R}_L > \tilde{R}_H$.
- In (one) equilibrium:

$$\begin{aligned} U(\frac{w_H}{p}(T-\tilde{R}_H), \tilde{R}_H, N\tilde{R}_H) &= U(\frac{w_H}{p}(T-R'_H-\kappa), R \\ U(\frac{w_L}{p}(T-R'_L), R'_L, (N-1)\,\tilde{R}_H + R'_L) &< U(\frac{w_L}{p}(T-\tilde{R}_L-\kappa), \tilde{R}_L \end{aligned}$$

where R'_i is the religious effort of wage type i if he joins club $k \neq i$.

Strategy of Manipulating Conflict

- What is the strategy of terror and how should targets of terror respond? What are the welfare implications of effective terror?
- We study "pure" logic of terrorism as information transmission and ask: "What is the strategic message of international terrorism?"
- "World War I was an unwanted spiral of hostility"..."World War II was not an unwanted spiral of hostility-it was a failure to deter Hitler's planned aggression."(Joseph Nye (2007)).
- Results depend critically on whether actions are *strategic substitutes* or *strategic complements*.

- According to *The Management of Savagery* (a document apparently composed by strategic thinkers within Al Qaeda) provoking U.S. will:
- "Force America to abandon its war against Islam by proxy and force it to attack directly so that the noble ones among the masses....will see that their fear of deposing the regimes because America is their protector is misplaced and that when they depose the regimes, they are capable of opposing America if it interferes." Abu Bakr Naji, *The Management of Savagery (*p. 24)
- Symmetrically, pacifists may try to convince moderates to become doves rather than hawks.
- Bertrand Russell founded the Campaign for Nuclear Disarmament (C.N.D.) which advocated unilateral nuclear disarmament. The slogan of this "ban the bomb" movement was "Better Red than Dead": "If no alternative remains except Communist domination or the extinction of the human race, the former alternative is the lesser of two evils."

- We allow an extremist to communicate information about the leader of their country to the other side. What is the effect of such cheap-talk on the probability of conflict? How does it depend on whether the extremist is a hawk or a dove?

Related Literature

- Bueno de Mesquita and Dickinson (2007) offer a model of provocation where one country can be "hard-line" or "soft-line". A hard-line government's preference for indiscriminate violence over negotiation is greater than a soft-line government's. Moderates in country B have to give up more in negotiations to a hard-line government than a soft-line government. (de Figueiredo and Weingast (2001) have a related contribution.)
- Kydd and Walter (2002) study "spoiling" where terrorists force an opponent to exit peace negotiations. The main idea is that a terror act signals that the leader in the terrorists' country is weak and/or a fanatic himself and cannot control the terrorists.
- These models are inspired by Spence signalling, where an informed sender sends a costly message to a receiver who then takes an action.
- We study pure cheap talk. To capture our key ideas, the strategic structure is necessarily more involved than in

Basic Model

- Two countries, A and B, with two leaders. Leaders can be interpreted as the pivotal decision-makers in the country, such as the median voter or dictator.
- Two actions: hawkish aggressive action (H) or dovish peaceful action (D). Cost of hawkish action for player i is c_i and payoffs for player i (the row player) are:

$$\begin{array}{ccc} & H & D \\ H & -c_i & \mu - c_i \\ D & -d & 0 \end{array} \tag{1}$$

  We assume $\mu > 0$ and $d > 0$. Action H may be an act of war, a vote for a hawkish political party or support for a hawkish faction. Action D is the reverse.
- The game has *strategic complements* if $d > \mu$ and *strategic substitutes* if $d < \mu$. Strategic complements (substitutes) captures the logic of escalation (deterrence).

- However, we are concerned with the following idea: Even a small initial fear may create an escalating cycle of fear that spirals out of control leading to mutual armaments. This was the basic insight of Schelling:

  "If I go downstairs to investigate a noise at night, with a gun in my hand, and find myself face to face with a burglar who has a gun in his hand, there is a danger of an outcome that neither of us desires. Even if he prefers to leave quietly, and I wish him to, there is a danger that he may think I want to shoot, and shoot first. Worse, there is danger that he may think that I think he wants to shoot. Or he may think that I think he thinks I want to shoot. And so on."

- Corcyraeans goaded Athens to go to war with Sparta: "Some of you may think there is no immediate danger of war. Those who think along these lines are deceiving themselves; they do not see the facts that Sparta is *frightened of you* and wants war" (our emphasis, Thucydides *History of the Peloponnesian War*, p. 55, 1.33).

- Player $i \in \{A, B\}$ has a type $c_i \in [\underline{c}, \overline{c}]$, $F'(c) > 0$ for all c.
- *Dominant strategy hawk*: H is a dominant strategy ($\mu \geq c_i$ and $d \geq c_i$).
- *Dominant strategy dove*: D is a dominant strategy ($\mu \leq c_i$ and $d \leq c_i$).
- *Coordination type:* H is a best response to H and D a best response to D ($\mu \leq c_i \leq d$).
- *Opportunistic type:* D is a best response to H and H a best response to D ($d \leq c_i \leq \mu$).
- Coordination types exist only with strategic complements, opportunistic types only with strategic substitutes.

	H	D
H	$-c_i$	$\mu - c_i$
D	$-d$	0

Assumption 1 Dominant strategy types of both kinds have positive probability: (1) If the game has strategic complements then $\underline{c} < \mu < d < \overline{c}$. (2) If the game has strategic substitutes then $\underline{c} < d < \mu < \overline{c}$.

Assumption 2 says that there is "enough uncertainty".

Assumption 2 $F'(c) < |\frac{1}{d-\mu}|$ for all $c \in [\underline{c}, \overline{c}]$.

(With a uniform distribution, Assumption 1 implies Assumption 2.)

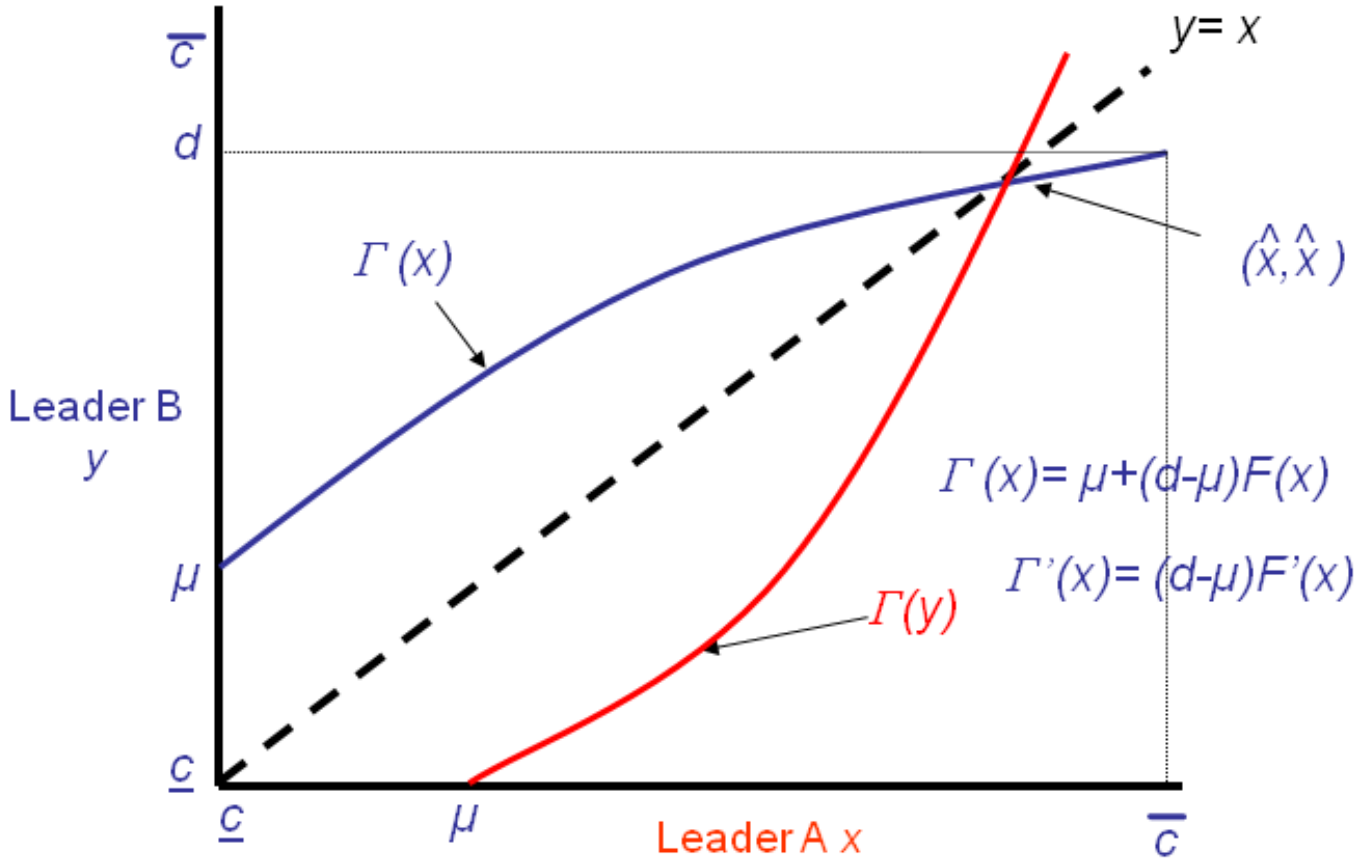
- Aggressive *dominant strategy hawks* play H regardless of the opponent's actions. Let the probability of these types be ε. But this triggers a multiplier effect.
- Some fraction $\delta > 0$ of all types are not dominant strategy hawks but prefer to play H when the opponent arms with at least probability ε. These "almost dominant strategy hawks" will play H as they know the dominant strategy hawks will do so.
- But then, all "almost-almost dominant strategy hawks" that prefer to play H when the opponent plays H with at least probability $\varepsilon + \delta$ will also arm, etc. The contagion takes hold.
- Similar arguments can be made for a cycle of "deterrence by fear" to create a unique equilibrium in the case of Strategic Substitutes.

Theorem

The conflict game has a unique Bayesian Nash equilibrium: Player i plays H iff $c_i \leq \hat{x}$.

- "Best response function" defined using cutoff strategies is upward (downward) sloping if actions are strategic complements (substitutes). In either case, a well-known sufficient condition for uniqueness is that best-response functions have slope strictly less than one in absolute value (see Vives's IO book). By Assumption 2, $1 > \Gamma'(x) > 0$ if $d > \mu$ and $-1 < \Gamma'(x) < 0$ if $d < \mu$. Hence, the best-response functions can cross at most once and there is a unique equilibrium.
- Technical aside: In global games, types are (highly) correlated rather than independent. Same kind of argument can be applied there. See Baliga and Sjöström, "*The Logic of Mutual Fear..*" (2008)

Figure 1. Strategic Complements:
Uninformative Equilibrium



Cheap Talk

- Add a third player, player E, the leader of an extremist group in country A. His payoff function is similar to player A's, with one exception: player E's cost type c_E differs from player A's cost type c_A.
- Two possibilities. Player E is a *hawkish extremist* ("terrorist") if $c_E < 0$. Player E is a *dovish extremist* ("pacifist") if $c_E > d + \mu$. His true type is commonly known.
- The hawkish extremist always wants player A to choose H. The dovish extremist always wants player A to choose D. Both want player B to choose D.

- *Player E knows c_A*. A terrorist or pacifist leader might know how likely it is that his extremist group will be able to influence the leader of his country. However, player E and A do not know c_B.

- Naji (p. 20): "[N]ote that the economic weakness resulting from the burdens of war or from aiming blows of vexation (al-nikāya) directly toward the economy is the most important element of cultural annihilation since it threatens the opulence and (worldly) pleasures which those societies thirst for. Then competition for these things begins after they grow scarce due to the weakness of the economy. Likewise, social iniquities rise to the surface on account of the economic stagnation, which ignites political opposition and disunity among the (various) sectors of society in the central country."
- Russell quote already motivated pacifist preferences.

- As $\mu > 0$ and $d > 0$, extremist wants player B to play D, whatever action player A chooses himself:

$$\begin{array}{cc} H & D \\ H \quad -c_i & \mu - c_i \\ D \quad -d & 0 \end{array}$$

- Aumann (1990) suggested that coordination on the efficient Nash equilibrium might then be hard.

Time Line

1. The cost type c_i is determined for each player $i \in \{A, B\}$. Players A and E learn c_A. Player B learns c_B.
2. Player E sends a (publicly observed) cheap-talk message $m \in M$.
3. Players A and B simultaneously choose H or D.

Cheap-talk is *effective* if there is a positive measure of types that choose different actions at time 3 than they would have done in the unique communication-free equilibrium. A PBE with effective cheap-talk is a *communication equilibrium*. For cheap-talk to be effective, player E's message must reveal some information about player A's type.

Monotonicity: for any message $m \in M$, there is a cut-off $c_j(m)$ such that if player j hears message m, then he chooses H if and only if $c_j \leq c_j(m)$.

Lemma. In a communication equilibrium, it is without loss of generality to assume $M = \{m_0, m_1\}$, where $c_B(m_1) > c_B(m_0)$. Player B is more likely to play H after m_1 than after m_0.

Cheap talk equilibria: Strategic Complements

Proposition. Doves can't Communicate Effectively.
Intuition: (Aumann intuition) With strategic complements, the message m_0 which makes player B more likely to play D must also make player A more likely to play D. But the dovish extremist will send m_0 even when his player A is a dominant strategy hawk so separation is impossible.

Proposition. If player E is a hawkish extremist and the game has strategic complements, then there exists a communication equilibrium. The hawkish extremist E uses cheap-talk to increase the risk of conflict above the level of the communication-free equilibrium. All types of players A and B are made worse off by this. If $F'(c) < \frac{1-F(\Gamma(d))}{d-\mu}$ for all $c \in (\underline{c}, \overline{c})$ then the communication equilibrium is unique.

- Recall that $M = \{m_0, m_1\}$. Interpret message m_1 as "terrorism". Terrorism occurs when c_A is an intermediate range (player A is a coordination type). In communication-free equilibrium, these types choose D. Terrorism causes them to switch to H for sure. Terrorism also makes player B more likely to choose H.
- If either c_A is very small or very large, then terrorism is counter-productive, because player A is not responsive to it. Because terrorism only occurs for intermediate values of c_A, it is an informative message.
- [Carlo Pisacane's] *"propaganda of the deed..recognized the utility of terrorism to a deliver a message to an audience other than the target and draw attention and support to a cause"*

- Interpret message m_0 as "no terrorism".
- "Curious incident of the dog in the night-time" (Conan Doyle): the terrorist in country A knows leader A's type. When the terrorist does not trigger a terror act, it can be because leader A is known to be a sympathizer with preferences aligned with the terrorist. After all, the terrorist "barks" when leader A is a weak type who plays D in the uninformative equilibrium.
- Hence, a terrorist who does not bark signals a greater likelihood that leader A is actually facing a dominant strategy hawk who plays H. This increases the incentive of leader B to play H and the logic of the reciprocal fear of surprise attack then implies the continuation equilibrium is more aggressive than the uninformative equilibrium.
- Separation of some types via terror acts triggers *greater* escalation.
- Overall, welfare goes down for all types of leader A and B relative to the uninformative equilibrium. It goes up for terrorists in some states of the world.

Strategic substitutes

- Many results are simply reversals of complements case.
- Only pacifists can speak informatively in equilibrium. They stage a peace protest when their leader is a strong opportunistic type who plays H in the uninformative equilibrium. Then, leader B plays H unless he is a dominant strategy dove and leader A backs off and plays D.
- When there is no peace protest, leader B learns there are no strong opportunistic types and becomes more aggressive. Leader A backs off and plays D more than in the uninformative equilibrium.
- It is not possible to determine if conflict goes up or down as leader B becomes more aggressive and leader A more dovish. But net effect can imply that pacifist action reduces probability of peace (D, D).
- The informative equilibrium has the "better red than dead" property: probability of leader B playing H and leader A playing D increases.

Strategic Effects of Ex Ante Investment

- Player B can make a publicly observed ex ante investment which increases his country's military capability.
- Suppose no extremist exists. With strategic substitutes, there is a incentive to over-invest in offensive capability (increase μ) in order to intimidate the opponent and force him to back down: *Top Dog strategy*. With strategic complements, there is a incentive to over-invest in defensive capability (reduce d) in order to reassure the opponent that one is unlikely to attack out of fear: *Fat cat strategy*.

- Re-introduce the extremist. With strategic complements, B's optimal strategy is still Fat Cat, making oneself look less threatening, same as before.
- With strategic substitutes, in the presence of a dovish extremist, B's optimal strategy is *also* Fat Cat. Thus, the presence of a dovish extremist dramatically changes the strategic effects.
- Intuition: the dovish extremist is, in a sense, an "ally" of player B, because peace protests make player A back down. In this case, Top Dog strategy can backfire for player B: by investing in offensive capacity, player B alarms the pacifist, who organizes fewer peace protests, which makes player B worse off.

Conclusion

- Hawkish extremists are either bad for peace (when actions are strategic complements) or irrelevant (when actions are strategic substitutes). Dovish extremists are either irrelevant (strategic complements) or have an ambiguous impact because they make one country more aggressive while the other backs down. In all cases, informative cheap-talk has a non-convex structure: it identifies a subset of moderate (intermediate) decision makers.

- Baliga and Sjöström (2004) showed that decision makers who talk to each other can identify "tough" moderates. Those types would have chosen H in the communication-free equilibrium, but after talking to each other feel safe to coordinate on D. Thus, communication makes them better off. In contrast, a hawkish extremist's message identifies "weak" moderates, who would have chosen D in the communication-free equilibrium, but now coordinate on H instead. Because the hawkish extremist creates conflict, players A and B have a common interest in preventing the hawkish extremist from communicating.
- The case of strategic substitutes was not considered by Baliga and Sjöström (2004). Here, we find that a dovish extremist's messages identifies "tough" moderates. Those types would have chosen H in the communication-free equilibrium, but now they back down and choose D. By strategic substitutes, the opponent then becomes more likely to choose H. Player B benefits from the activity of the pacifist, but player A would like to suppress it.

Torture

- A terrorist attack is planned for a major holiday, a few weeks from now. A suspect with potential intelligence about the impending attack awaits interrogation.
- One possibility: Suspect is *innocent* or *uninformed*.
- Other possibility: Suspect is *informed*.
- In this situation, suppose *torture* is the only instrument available to obtain information.

- Torture is costly and/or abhorrent to society. But given the expected value of information, the "best" option may be to torture to try to extract information from the suspect.
- Michael Walzer (1973): "[A] politician...is asked to authorize the torture of a captured rebel leader who knows or probably knows the location of a number of bombs hidden in apartment buildings around the city, set to go off within the next twenty-four hours. He orders the man tortured, convinced that he must do so for the sake of the people who might otherwise die in the explosions..."
- Liz Cheney : "Mr. President, in a ticking time-bomb scenario, with American lives at stake, are you really unwilling to subject a terrorist to enhanced interrogation to get information that would prevent an attack?"

Strategic Dilemmas of Torture

The most effective use of torture requires

- A commitment to torture a victim who is known to be uninformed.
- A promise not to continue torturing a victim who is known to be informed.

These are inconsistent with the rationale for torture: a victim who is known to be guilty must be threatened with torture till he gives up everything. A victim who is uninformed should not be tortured. We analyze the value of torture when the credibility of threats and promises is taken into account.

- Main questions: What is the optimal policy for the principal? How does the ability to torture for a long time or more frequently affect the optimal scheme, *ceteris paribus*? How to "enhanced interrogation techniques" affect the optimal scheme, *ceteris paribus*? Etc., etc...

Comparative Statics:

1. Value of time disappears for the principal: It is never optimal to torture more than a limited amount of time however much information the agent has.
2. If the principal has frequent opportunities to torture in a given time period, the value of torture approaches zero.
3. Enhanced interrogation techniques can *reduce* the principal's welfare. A ban on enhanced interrogation techniques by reducing available torture instruments can increase welfare.

Related Literature

- When the principal discovers the agent is uninformed, he has the incentive to stop torturing (renegotiation and Coase conjecture: Hölmström and Myerson (1983), Dewatripont (1989), Fudenberg and Tirole (1983), Sobel and Takahashi (1983), Gul, Sonnenschein and Wilson (1986) and Hart and Tirole (1988)).
- If the principal discovers the agent is informed, he has the incentive to extract more information ("ratchet effect": Freixas, Guesnerie and Tirole (1985) and Laffont and Tirole (1988)).

Torture

- Principal and agent (victim, suspect).
- The suspect is either *informed* with a quantity x of perfectly divisible, verifiable (i.e. "hard") information or *uninformed.*
- Let $\mu_0 \in (0,1)$ be the prior probability that the suspect is informed.
- A terrorist attack is known to be planned time T after the principal receives the suspect.
- Torture imposes a flow cost of $\Delta > 0$ on the suspect.
- It imposes a flow cost $c > 0$ on the principal.

Payoffs

- If the principal tortures the agent for time $t \leq T$ and the informed suspect reveals $y \leq x$ to the principal, payoffs are

$$\begin{aligned} \text{Agent:} \quad & -y - \Delta t \\ \text{Principal} \quad : \quad & y - ct \end{aligned}$$

Full Commitment: Torture as Mechanism Design Problem

- The principal can commit to a torture schedule.
- There is no individual rationality constraint.

Optimal Mechanism

- The principal demands information $y \leq x$ from the suspect. If he does not reveal this amount of information, he tortures him for $t(y) \leq T$ periods where $t(y) = \frac{y}{\Delta}$.
- The principal's payoff is

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

and we have the following solution:

- **Theorem:** *At the full commitment solution, if* $\mu_0\Delta - (1-\mu_0)\, c \geq 0$, *the principal demands information* $\min\{x, T\Delta\}$ *and inflicts torture for* $\min\{\frac{x}{\Delta}, T\}$ *periods at all other levels of information revelation. If* $\mu_0\Delta - (1-\mu_0)\, c < 0$, *the principal does not demand any information and does not torture at all.*

Torturing the Innocent

In the optimal mechanism,

- only the uninformed victim suffers torture.
- torture continues when it is certain it will yield no more information.

These features reflect the principal's ability to commit to a torture plan.

Model with Limited Commitment

- For convenience, we assume time is divided into intervals of length 1 and there are T periods total.
- We measure time in reverse, so "period k" means that there are k periods remaining.
- Define $\bar{k}$ by $(\bar{k}+1)\,\Delta \geq x \geq \bar{k}\Delta$. If $k \leq \bar{k}$ we say the game is in the *ticking time bomb phase*.
- Principal can only commit to torture for a single period.
- Principal can demand a quantity of information $y \geq 0$ and commit to suspend torture in the given period if it is given.
- At the end of each period the principal decides whether to cease torturing or continue, but he cannot commit in advance to cease or continue.

Themes

- Spilling your guts.
- Torture must continue for as long as informed type talks in equilibrium.
- This two features allow us to determine value of torture.

Spill Your Guts

- Suppose agent stay quiet till period 2 and the concedes $y > 0$ in period 2.
- Principal knows agent is informed and the game has complete information.
- Principal demands $y = \Delta$ in period 1 and threatens torture if any less is offered.
- Threat is never carried out in equilibrium and principal gets payoff Δ and agent gets $-\Delta$.
- This is true more broadly:

Lemma

(**Spill your guts lemma**) *In any equilibrium, at the beginning of the complete information continuation game with k periods remaining and a quantity $\tilde{x}$ of information yet to be revealed, the suspect's payoff is*

$$-\min\left\{\tilde{x}, k\Delta\right\}.$$

Period One

- Suppose $x > 3\Delta$ and $T = 3$ so we are in ticking time bomb phase.
- Suppose principal's posterior is μ and let

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

  - A demand $y > \Delta$ is rejected by the agent and a demand $y < \Delta$ is dominated for the principal.
- Let μ_1^* be defined by

$$\mu_1^*\Delta - (1-\mu_1^*)c = 0.$$

- If $\mu_1 \geq \mu_1^*$, it is optimal for principal to demand $y = \Delta$ and torture if any less information is given. Otherwise, set $y = 0$.

Period Two

- Spill your guts lemma implication: If agent concedes $y > 0$ in period 2, he knows he will get payoff $-y - \Delta$.
- To maintain a two period torture regime, the principal must be willing to torture in period 1 if suspect does not give up information in period 2. Let q be the probability that informed type concedes information in period 2 and let μ be principal's belief.
- Let $q_2(\mu)$ be the probability of concession in *period 2* that makes the principal indifferent between torturing and not in *period 1*:

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

- If $q > q_2(\mu)$, $\mu_1 < \mu_1^*$ and the principal's posterior in period 1 implies he will *not* torture. But then agent should never concede in period 2. Hence, we must have $q \leq q_2(\mu)$.

Period Two

- If $q < q_2(\mu)$, $\mu_1 > \mu_1^*$ and the principal will torture with probability 1 in period 1. If principal demands $y < \Delta$ in period 2, agent receives

$$-y - \Delta$$

if he concedes in period 2 and

$$-2\Delta$$

if he resists. So, he will concede with probability 1, contradiction.

- We must have $q = q_2(\mu)$ and $y = \Delta$ in equilibrium.

Period Two

- Principal's payoff with posterior μ is

$$V^2(\mu) = q_2(\mu)\mu 2\Delta + (1 - q_2(\mu)\mu)\left(V^1(\mu_1^*) - c\right)$$

or

$$\begin{aligned} & V^2(\mu) = q_2(\mu)\mu 2\Delta - c\left(1 - q_2(\mu)\mu\right) \\ & + (1 - q_2(\mu)\mu)\left(\frac{\mu(1 - q_2(\mu))}{1 - \mu q_2(\mu)}\Delta - (1 - \frac{\mu(1 - q_2(\mu))}{1 - \mu q_2(\mu)})c\right) \\ = \;& q_2(\mu)\mu 2\Delta - c\left(1 - q_2(\mu)\mu\right) + \Delta\mu\left(1 - q_2(\mu)\right) - c(1 - \mu) \\ = \;& q_2(\mu)\mu\Delta - c\left(1 - q_2(\mu)\mu\right) + V^1(\mu) \\ = \;& q_2(\mu)\mu\Delta - c\mu\left(1 - q_2(\mu)\right) - (1 - \mu)c + V^1(\mu). \end{aligned}$$

Period Two

- Since the informed type only talks with positive probability in period 2, the "marginal benefit" of torture is lower and the "marginal cost" is higher than in period 1:

$$V^2(\mu) - V^1(\mu) = q_2(\mu)\mu\Delta - c\mu\left(1 - q_2(\mu)\right) - (1-\mu)c. \tag{2}$$

  Per period benefits of torture decline and costs go up as torture increases in length as informed pretend to be uninformed.

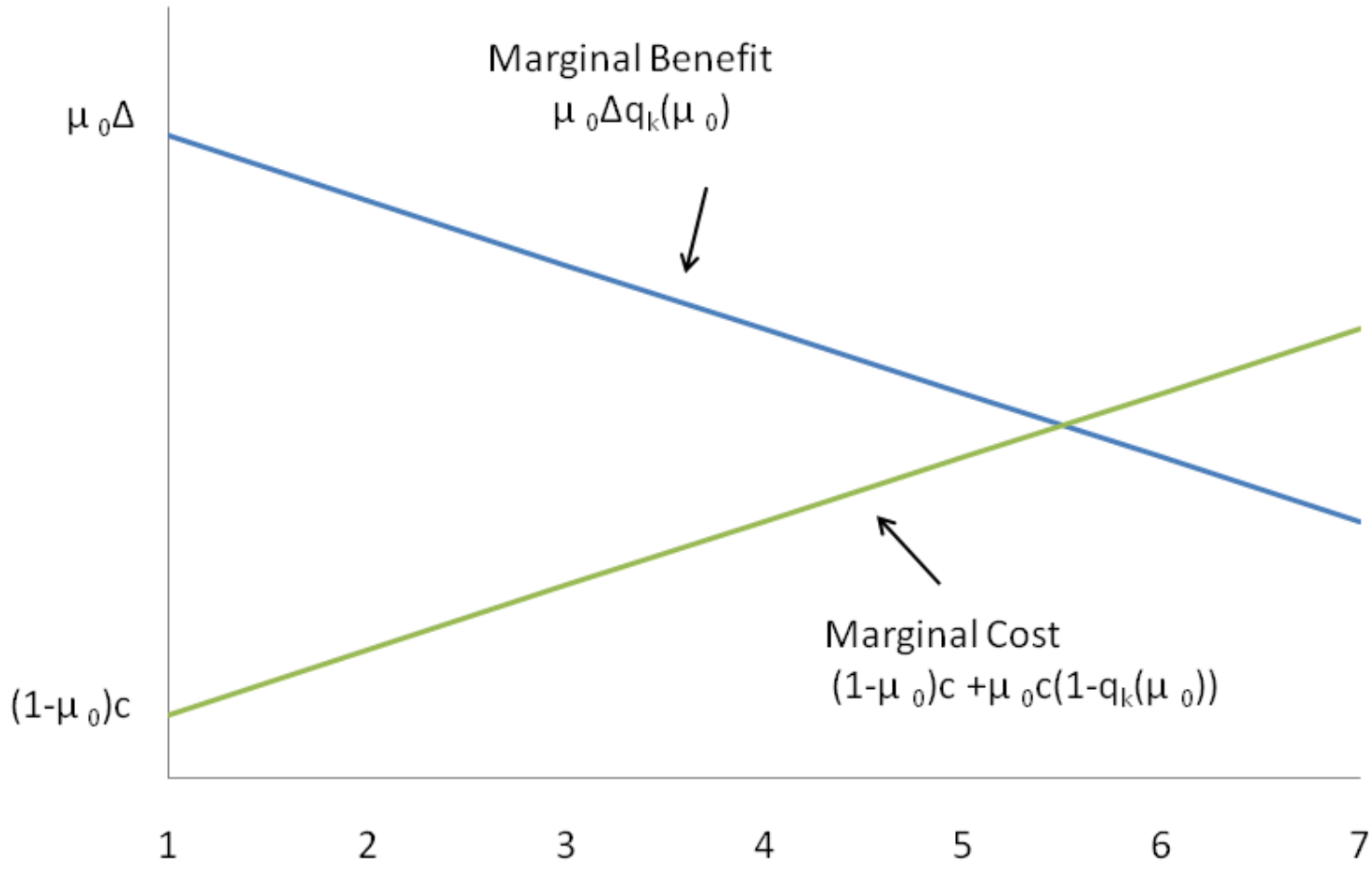
- The principal tortures for two periods rather than one iff (2) is positive.

Period Three

- An informed type who begins to concede in period 3 knows he will spill his guts for three periods total. If torture begins in period 3, principal must be willing to torture in periods 2 and 1 if suspect does not give up information. Hence, $q_3(\mu)$ must make principal indifferent between torturing in period 2 or waiting till period 1. It leads to belief μ_2^* such that

$$\begin{aligned} V^2(\mu_2^*) - V^1(\mu_2^*) &= q_2(\mu_2^*)\mu_2^*\Delta - c\,(1-\mu_2^*) - c\mu_2^*(1-q_2(\mu_2^*)) \\ &= 0. \end{aligned}$$

- *Key Point:* Notice $\mu_2^* > \mu_1^*$ as $q_2(\mu_2^*) < 1$ and also therefore $q_2(\mu_0) > q_3(\mu_0)$.
- *The longer is the torture regimen, the slower is the rate of confession and the higher must be the principal's beliefs to implement the regimen.*


Marginal Benefit
$\mu_0 \Delta q_k(\mu_0)$
$\mu_0 \Delta$
Marginal Cost
$(1-\mu_0)c + \mu_0 c(1-q_k(\mu_0))$
$(1-\mu_0)c$
1
2
3
4
5
6
7

General Case

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

$$V^k(\mu) = \mu q_k(\mu) \min\{x, k\Delta\} + (1 - \mu q_k(\mu)) \left[V^{k-1}(\mu_{k-1}^*) - c \right]. \tag{3}$$

$$V^k(\mu_k^*) = V^{k-1}(\mu_k^*) \tag{4}$$

$$\frac{\mu(1 - q_k(\mu))}{1 - \mu q_k(\mu)} = \mu_{k-1}^*. \tag{5}$$

Theorem

The unique equilibrium payoff for the principal is

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

Remark: It is never optimal to torture more than one period outside the ticking time bomb phase.

Limits of Torture

- Increase number of periods T available to torture as well as x. Study what happens to value of torture.
- It increases value of torture if principal can fully commit. But what if he cannot?
- *Key Idea:* The probability the agent is informed must be higher, the longer the planned torture regimen. But it may be impossible to maintain such a long regime given the prior μ_0.

Theorem

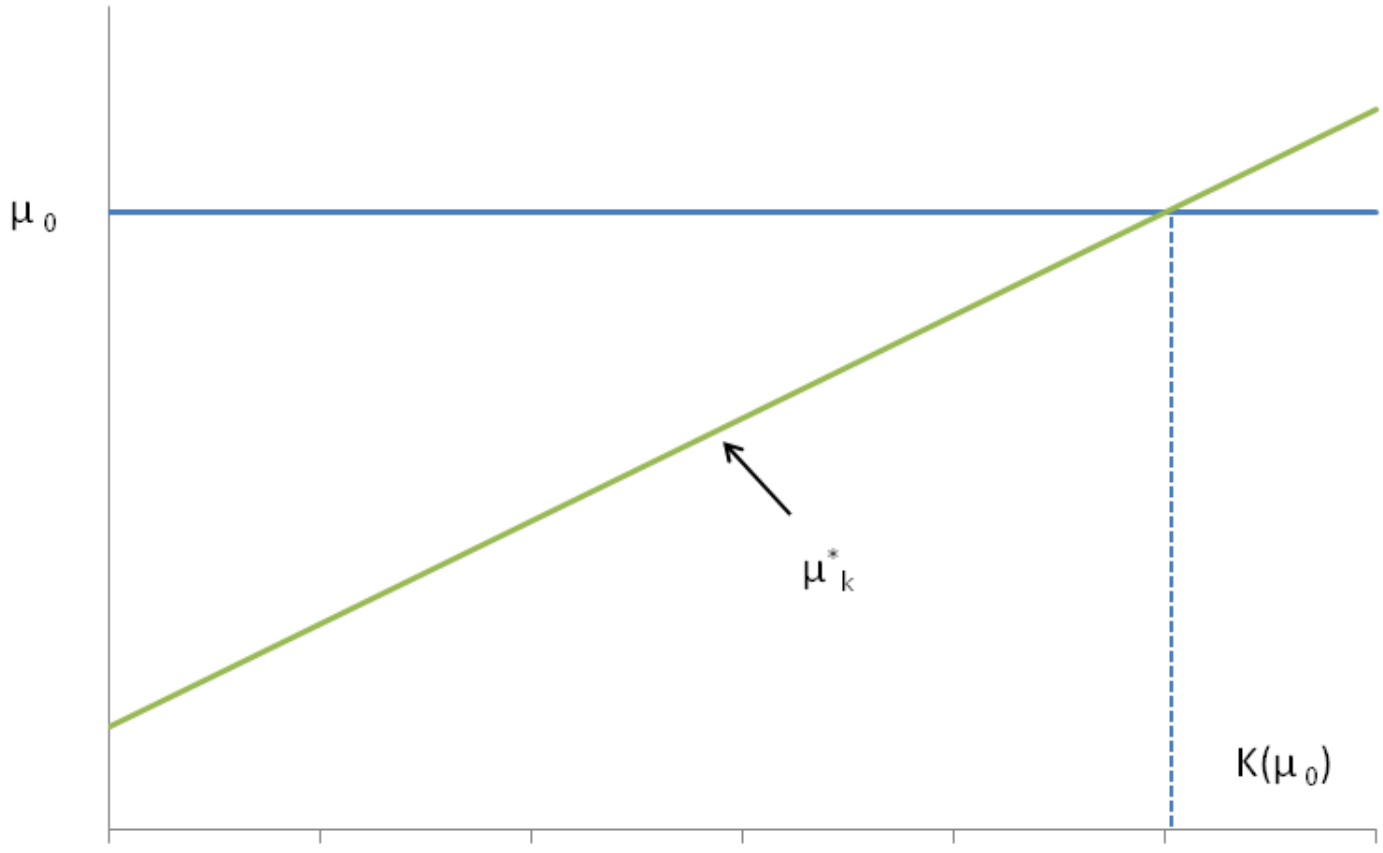
Fix the prior μ_0 and define let $K(\mu_0)$ to be the largest k such that the sum

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

is no larger than μ_0.

1. *Regardless of the value of x, the principal tortures for at most $K(\mu_0)$ periods.*
2. *Regardless of the value of x, the principal's payoff is less than* $\max_{k\leq K(\mu_0)} V^k(\mu_0)$.

Note that for any given μ_0, the displayed sum converges to infinity in k and therefore $K(\mu_0)$ is finite for any μ_0.
Laws against indefinite detention do not compromise ability to extract information.


μ_0
μ^*_k
$K(\mu_0)$

Frequency of Torture

- Reduce the principal's commitment power while keeping everything else equal, e.g. x, total *physical* time available for torture.
- If period has length l, costs of torture are $l\Delta$ and lc. If principal can fully commit this has no effect: principal threatens torture for same length of physical time as before.
- We can reformulate limited commitment problem so

$$\begin{aligned} V_l^k(\mu) &= \mu q_k(\mu|l) kl\Delta + (1 - \mu q_k(\mu|l)) \left[V_l^{k-1}(\mu^*_{k-1,l}) - lc \right] \\ &= lV^k(\mu) \end{aligned}$$

  l times the payoff when period length is 1.
- From "limits of torture" section it follows that payoff is bounded above by $l \max_{k \leq K(\mu_0)} V^k(\mu_0)$.

Enhanced Interrogation Techniques

- Suppose $T = 2$ and consider two technologies: "sleep deprivation", which delivers flow costs Δ and c and "water boarding", which delivers costs Δ' and c'. Water boarding has benefits because $\Delta' > \Delta$ but comes at a cost, $c'/\Delta' > c/\Delta$.
- Depending on the parameters (e.g. low μ_0), it is optimal to use sleep deprivation for two periods if the principal can make a once-and-for-all commitment to stick with the same technology.
- But the principal can switch technologies in mid-stream. If agent talks in period 2, principal will switch to waterboarding to extract Δ' at no cost. The agent's payoff is $-\Delta - \Delta'$ if he talks in period 2.
- This causes the equilibrium to unravel: If the agent stay quiet in period 2, the principal uses sleep deprivation in period 1 and agent's payoff is -2Δ.
- In equilibrium, the principal tortures for one period with waterboarding.
- This is another "ratchet effect" in our model.

Problems with Commitment: Delegation

- Suppose the principal can utilize a sadistic specialist who *benefits* $c' < 0$ from using torture.
- This eliminates the commitment problem of torturing innocent and hence increases information revelation.
- But this comes at the cost of too much torture: The specialist will torture the agent in all periods when he is not extracting information.
- An optimally timed employment and dismissal of specialist seems to fix this problem.
- But this reveals the fundamental problem with the delegation scenario: As torture is carried out in secret because of its very nature, the principal can terminate the specialist as soon as the victim does not reveal information.
- This generates the commitment problems we study yet again.

Conclusion

- Torture without commitment has weak power to extract information and becomes close to useless the more time and the less commitment you have.
- A purported Al Qaeda manual recommends to the captured terrorist: "The brother may think that by giving a little information he can avoid harm and torture. However, the opposite is true. The torture and harm would intensify to obtain additional information, and that cycle would repeat. Thus, the brother should be patient, resistant, silent, and prayerful to Allah, especially if the security apparatus knows little about him."