Honest Threats: The Interaction of Reputation and Political Institutions in International Crises

Alexandra Guisinger
Department of Political Science
Yale University
New Haven, CT 06511

Alastair Smith
Department of Political Science
Yale University
New Haven, CT 06511

October 16, 2001

The document contains the mathematical appendix for "Honest Threats: The Interaction of Reputation and Political Institutions in International Crises." At the editor's request it is provided here, as a webpage, rather than being printed with the main article.

APPENDIX

In each period a crisis occurs between nations A and B. v_A^t and v_B^t are randomly and independently drawn from the distributions $F_A(.)$ and $F_B(.)$, respectively. At the beginning of each period each player learns the value of value for the issue, v_i $(i = A, B)$. Player $j \neq i$ does not learn the value v_i^t, but does know the distribution from which it was drawn. Beyond the period by period evaluation of the issue, the structure of the game and all other payoff are complete information to all players. Payoffs in future periods are discounted by a common factor δ.

Prior to the playing the crisis game, B announces a declaration of foreign policy. Since this announcement is a cheap talk signal and subsequently B choices are binary, we only consider two messages (Crawford and Sobel 1982). Thus we limit the set of messages to $M = R \cup S$, where the messages R and S are implicitly understood to mean resist and surrender, respectively.

Let h^t represent the history of play during all subgames prior to period t and let H^t be the set of all possible histories. For much of what follows we only utilize certain features of past play , we partition H^t into $cheat^t$ and $honest^t$, where $cheat^t$ represents the subset of H^t in which B has never stated that it would (R) resist but failed to resist. If B has always resisted if challenged when it said it would resist (R) then $h^t \in honest^t$.

Let $s_A^t(h^t, M, v_A)$ represent the probability that Country A attacks Country B given the message M, the history of previous play h^t and A's current period valuation of the

issue v_A^t. Let $s_B^t(h^t, M, v_B^t)$ represent the probability that B resists given it stated the message M, the history of previous play h^t and B's current valuation of the issue v_B^t. $\sigma_B^t(M, h^t, v_B^t)$ is B's message sending strategy: the probability that B sends message M, given history h^t and evaluation v_B^t.

Expected value of uninformative equilibria

We start by calculating B's *ex ante* payoff from playing the stage game. By *ex ante*, we mean prior to learning type. $E[U_B(crisis)] = \Pr(SQ)E[v_B] + \Pr(Aattacks)$ $\Big(\Pr(v_B < \underline{v}_B)0 + \int_{\underline{v}_B}^1 f_B(v_B)((1-p)v_B - k_B)dv_B\Big)$

$= F_A(\underline{v}_A)E[v_B] + (1 - F_A(\underline{v}_A)) \int_{\underline{v}_B}^1 f_B(v_B)((1-p)v_B - k_B)dv_B$, where $f_B(v_B)$ is the probability density of B types ($f_B(v_B) = \frac{dF_B(v_B)}{dv_B}$). For the special case of the uniform distribution this reduces to $(1-\alpha)\frac{1}{2} + \alpha\beta\left((1-p)\frac{1+\underline{v}_B}{2} - k_B\right)$, where $\underline{v}_B = \frac{k_B}{1-p}$, $\beta = 1 - \frac{k_B}{1-p}$, $\alpha = 1 - (\frac{\beta k_A}{1-\beta+p\beta}) = 1 - k_A\frac{1-p-k_B}{(1-p)(p+k_B)}$. So $E[U_B(crisis)] = (1 - \left(1 - k_A\frac{1-p-k_B}{(1-p)(p+k_B)}\right))\frac{1}{2} + \left(1 - k_A\frac{1-p-k_B}{(1-p)(p+k_B)}\right)\left(1 - \frac{k_B}{1-p}\right)\left((1-p)\frac{1+\frac{k_B}{1-p}}{2} - k_B\right)$. Given this per period payoff, the continuation value for playing the non-informative strategy in all future periods is $W_u = \sum_{\tau=0}^{\infty}\delta^\tau E[U_B(crisis)] = \frac{1}{1-\delta}E[U_B(crisis)]$.

Informative equilibria

With respect to informative equilibria we examine the class of strategies where whether A conditions its response to B's message depends upon the history of past play. In particular we look at trigger strategies where A ignores B message if B has ever lied in any previous period: $s_A^t(R, h^t, v_A^t) = s_A^t(S, h^t, v_A^t)$ for all t, for all v_A^t, if $h^t \in Cheat^t$ and $s_A^t(R, h^t, v_A^t) \leq s_A^t(S, h^t, v_A^t)$ for all t, for some v_A^t, if $h^t \in Honest^t$.

Country-Contingent Reputation Strategy is as follows

1) If $h^t \in cheat^t$ then $\sigma_B^t(M, h^t, v_B^t) = \sigma_B^t(M, h^t, v_B^{tt})$ for all v_B^t,v_B^{tt}, for all M.

If $h^t \in honest^t$ then $\sigma_B^t(R, h^t, v_B^t) = \begin{cases} 1 & if \quad v_B^t \geq v_B^\ddagger \\ 0 & if \quad v_B^t < v_B^\ddagger \end{cases}$

2) If $h^t \in cheat^t$ then $s_B^t(h^t, M, v_B^t) = \begin{cases} 1 & if \quad v_B^t \geq \underline{v}_B = \frac{k_B}{1-p} \\ 0 & if \quad v_B^t < \underline{v}_B = \frac{k_B}{1-p} \end{cases}$, for all M.

If $h^t \in honest^t$ then $s_B^t(h^t, R, v_B^t) = \begin{cases} 1 & if \quad v_B^t \geq \widehat{v_B} = \frac{k_B - \delta(W_h - W_c)}{1-p} \\ 0 & if \quad v_B^t < \widehat{v_B} = \frac{k_B - \delta(W_h - W_c)}{1-p} \end{cases}$,

and $s_B^t(h^t, S, v_B^t) = \begin{cases} 1 & if \quad v_B^t \geq \underline{v}_B = \frac{k_B}{1-p} \\ 0 & if \quad v_B^t < \underline{v}_B = \frac{k_B}{1-p} \end{cases}$

3) If $h^t \in cheat^t$ then $s_A^t(h^t, M, v_A^t) = \begin{cases} 1 & if \quad v_A^t \geq \underline{v}_A = \beta \frac{k_A}{1-\beta+\beta p} \\ 0 & if \qquad v_A^t < \underline{v}_A \end{cases}$, where $\beta = 1 - F_B(\underline{v}_B) = 1 - (\frac{k_B}{1-p})$.

If $h^t \in Honest^t$ then $s_A^t(h^t, R, v_A^t) = \begin{cases} 1 & if \quad v_A^t \geq \widehat{v_A} = \beta(R) \frac{k_A}{1-\beta(R)+\beta(R)p} \\ 0 & if \qquad v_A^t < \widehat{v_A} \end{cases}$ and if $h^t \in Honest^t$ then $s_A^t(h^t, S, v_A^t) = 1$ for all v_A^t.

The terms $\beta(R)$, $v_B^\ddagger$, $\underline{v}_B$, $\widehat{v_B}$, $\underline{v}_A$, and $\widehat{v_A}$ are defined below.

If nations play the CCR strategy then we can characterize observable behavior as follows: Let β represent the probability that B resists given a dishonest reputation: $\beta = 1 - F_B(\frac{k_B}{1-p})$. The probability that A attacks given a dishonest reputation is $\alpha = \Pr(v_A \geq \underline{v}_A) = 1 - F_A(\underline{v}_A)$, where $\underline{v}_A = \frac{\beta k_A}{1-\beta+p\beta}$. If B has an honest reputation and send message R then the probability of resistance, $\beta(R) = \Pr(v_B^t \geq \widehat{v_B} | v_B^t \geq v_B^\ddagger) = \begin{matrix} 1 & if \quad \widehat{v_B} \leq v_B^\ddagger \\ \frac{1-F_B(\widehat{v_B})}{1-F_B(v_B^\ddagger)} & if \quad \widehat{v_B} > v_B^\ddagger \end{matrix}$, if B sends message S then the probability of resistance is $\beta(S) = \Pr(v_B^t \geq \underline{v}_B | v_B^t < v_B^\ddagger) = 0$. Given an honest reputation, the probability A attacks given message R is $\alpha(R) = 1 - F_A(\frac{\beta(R) k_A}{1-\beta(R)+p\beta(R)})$, and the probability of attack given message S is $\alpha(S) = 1$.

The continuation value represents the expected value of playing the infinitely repeated game under CCR. We let W_h and W_c represent the continuation values if B has an honest or a dishonest reputation, respectively.

Under CCR, if B has lost its reputation for honesty, then there is no informative signaling so B's per period payoff is identical to that of the non-informative case: $W_c = \sum_{\tau=0}^{\infty} \delta^\tau E[U_B(crisis)] = \frac{1}{1-\delta}(F_A(\underline{v}_A) E[v_B] + (1 - F_A(\underline{v}_A)) \int_{\underline{v}_B}^1 f_B(v_B)((1 -$

$p)v_B - k_B)dv_B)$, which for the special case of the uniform distribution this reduces to $\frac{1}{1-\delta}((k_A\frac{1-p-k_B}{(1-p)(p+k_B)})\frac{1}{2} + \left(1 - k_A\frac{1-p-k_B}{(1-p)(p+k_B)}\right)\left(1 - \frac{k_B}{1-p}\right)\left((1-p)\frac{1+\frac{k_B}{1-p}}{2} - k_B\right))$.

We now calculate the continuation value if B has an honest reputation. There are two cases: 1) All signals are fully informative and only types that will subsequently resist declare an intention to do so, and 2) signals are not fully informative and some types bluff, declaring an intention to resist but not actually resisting if attacked.

Case 1). If $v_B^\ddagger \geq \widehat{v_B}$ (i.e. all threats are credible) then

$W_h = F_B(v_B^\ddagger)\,(0 + \delta W_h) + \int_{v_B^\ddagger}^{\infty}\left(\delta W_h + \alpha(R)((1-p)v_B - k) + (1-\alpha(R))v_B\right) f_B(v_B)dv_B$.
For the special case of the uniform distribution, $W_h = \frac{1}{(1-\delta)}(1-v_B^\ddagger)(\frac{1+v_B^\ddagger}{2} + \alpha(R)(-p\frac{1+v_B^\ddagger}{2} - k_B))$, where $\alpha(R) = 1 - \frac{k_A}{p}$.

Case 2). Else if $\widehat{v_B} > v_B^\ddagger$ (i.e. some types bluff and are not prepared to carry out their threat) then

$W_h = F_B(v_B^\ddagger)\,(0 + \delta W_h) + \int_{v_B^\ddagger}^{\widehat{v_B}}\left((1-\alpha(R))\,(v_B + \delta W_h) + \alpha(R)\,(0 + \delta W_u)\right) f_B(v_B)dv_B$
$+ \int_{\widehat{v_B}}^{\infty}\left(\delta W_h + \alpha(R)((1-p)v_B - k) + (1-\alpha(R))v_B\right) f_B(v_B)dv_B$.

There are two cases of equilibria. Fully informative messages ($v_B^\ddagger \geq \widehat{v_B}$ and hence $\beta(R) = 1$) and partially informative equilibria ($v_B^\ddagger < \widehat{v_B}$ and hence $\beta(R) = \frac{1-F_B(\widehat{v_B})}{1-F_B(v_B^\ddagger)}$). In the main text we focus on the former case, a practice we will continue here.

Proposition: The Country-Contingent Reputation Strategy is a perfect Bayesian Equilibrium, where with an honest reputation messages are fully informative of B's intention to resist ($\beta(R) = 1$, and $\beta(S) = 0$), if $\delta \geq k_B\frac{1-\alpha(R)}{(1-\alpha(R)p)(W_h - W_c)}$.

Proof: i) We start by considering the eventually where B has cheated in a previous period. Suppose $h^t \in cheat^t$. Under this contingency, A never conditions its strategy on B's messages so all messages generate the same expected rewards. Hence, if $h^t \in cheat^t$ then $\sigma_B^t(M, h^t, v_B^t) = \sigma_B^t(M, h^t, v_B^{t\prime})$ for all v_B^t,$v_B^{t\prime}$, for all M is an optimal strategy. Given B's message sending strategy, messages are uninformative and hence A's optimal behavior is not conditioned by the message. Hence, A and B's conflict behavior are identical to those in the initially analyzed uninformative case.

ii) Next we consider crisis behavior under the contingency where B has always been honest in the past: $h^t \in honest^t$.

iia) Suppose $M = R$. Given that A has attacked if B surrenders then A will never believe any future messages and so B's future payoff are $0 + \delta W_c$. If B resists then her payoff is $(1-p)v_B - k_B + \delta W_h$. Hence B resists iff $v_B \geq \widehat{v_B} = \frac{k_B - \delta(W_h - W_c)}{1-p}$. If $v_B^\ddagger \geq \widehat{v_B}$, then by Bayes rule the probability B resists having sent message R is $\beta(R) = 1$. Alternatively, if $v_B^\ddagger < \widehat{v_B}$, then by Bayes rule the probability B resists having sent message R is $\beta(R) = \frac{1-F_B(\widehat{v_B})}{1-F_B(v_B^\ddagger)}$.

A's expected payoff for attacking is $(1-\beta(R))v_A + \beta(R)(pv_A - k_A)$. Without challenging. A's payoff is 0. Hence, A attacks iff $v_A \geq \widehat{v_A}$, where $\widehat{v_A} = \beta(R)\frac{k_A}{1-\beta(R)+\beta(R)p}$. Hence the probability that A attacks given message R is $\alpha(R) = (1 - F_A(\widehat{v_A}))$.

iib) Suppose $M = S$. Given that A has attacked if B surrenders its expected payoff is $0 + \delta W_h$. Alternatively, resisting give B a payoff of $(1-p)v_B - k_B + \delta W_h$. Hence $s_B^t(h^t, S, v_B^t) > 0$ iff $v_B^t \geq \underline{v}_B = \frac{k_B}{1-p}$. Let $\beta(S)$ be the probability that B resists having sent message S. Since $\underline{v}_B > v_B^\ddagger$, by Bayes rule $\beta(S) = 0$. If A attacks given message S then its expected payoff is $(1-\beta(S))\, v_A + \beta(S)(pv_A - k_A) = v_A$. Without challenging. A's payoff is 0. Hence, A always attacks: $\alpha(S) = 1$.

iii) We now consider B's message sending strategy. Having learned its type, if B sends message R, then B's expected payoff is $(1-\alpha(R))(v_B + \delta W_h) + \alpha(R)Z(v_B)$, where $Z(v_B) = MAX\{(1-p)v_B - k_B + \delta W_h, \delta W_c\}$. If B sends message S, then B's expected payoff is $0 + \delta W_h$. We let type $v_B^\ddagger$ be the type indifferent between these messages.

Case 1 ($v_B^\ddagger \geq \widehat{v_B}$): $(1-\alpha(R))(v_B + \delta W_h) + \alpha(R)((1-p)v_B - k_B + \delta W_h) = 0 + \delta W_h$ which implies $v_B^\ddagger = \alpha(R)\frac{k_B}{1-\alpha(R)p}$, where $\alpha(R) = 1 - F_A(\frac{k_A}{p})$.

Case 2 ($v_B^\ddagger < \widehat{v_B}$): $(1-\alpha(R))(v_B + \delta W_h) + \alpha(R)(\delta W_c) = \delta W_h$ which implies that $v_B^\ddagger = \delta\alpha(R)\frac{W_h - W_c}{1-\alpha(R)}$, where $\alpha(R) = 1 - F_A(\frac{\beta(R)k_A}{1-\beta(R)+p\beta(R)})$, $\beta(R) = \frac{1-F_B(\widehat{v_B})}{1-F_B(v_B^\ddagger)}$, $W_c = \frac{1}{1-\delta}(F_A(\underline{v}_A)E[v_B] + (1 - F_A(\underline{v}_A))\int_{\underline{v}_B}^1 f_B(v_B)((1-p)v_B - k_B)dv_B)$, and $W_h =$

$F_B(v_B^\ddagger)\,(0+\delta W_h) + \int_{v_B^\ddagger}^{\widehat{v_B}} \left((1-\alpha(R))\,(v_B+\delta W_h) + \alpha(R)\,(0+\delta W_u)\right) f_B(v_B)dv_B$
$+ \int_{\widehat{v_B}}^{\infty} \left(\delta W_h + \alpha(R)((1-p)v_B - k) + (1-\alpha(R))v_B\right) f_B(v_B)dv_B$. An explicit characterization of equilibria in this scenario requires simultaneously solving this series of equations.

Messages are completely informative of B's intention to resist in case 1: $(v_B^\ddagger \geq \widehat{v_B})$. This requires that $(v_B^\ddagger = \alpha(R)\frac{k_B}{1-\alpha(R)p} \geq \frac{k_B-\delta(W_h-W_c)}{1-p} = \widehat{v_B})$ is satisfied when $\delta \geq k_B\frac{1-\alpha(R)}{(1-\alpha(R)p)(W_h-W_c)}$. Hence CCR is a perfect Bayesian equilibrium. QED

Agent-Contingent Reputation Strategy

The ACR strategy is similar to the CCR. Hence rather than introduce additional notation, we let s_B^t and σ_B^t represent the strategy of incumbent leader B at time t. Once removed from office we assume leaders have a negligible probability of returning to office. Let $s_E^t(h^t)$ represent the probability with which citizens retain leaders given the incumbents history of past play. In contrast to the earlier case, an honest reputation, $h^t \in honest^t$, requires only that the incumbent has never failed to follow through on a threat, it does not requires that her predecessors were also honest. The Agent-Contingent Reputation Strategy is as follows:

1) If $h^t \in cheat^t$ then $\sigma_B^t(M,h^t,v_B^t) = \sigma_B^t(M,h^t,v_B^{t\prime})$ for all v_B^t,$v_B^{t\prime}$, for all M.

If $h^t \in honest^t$ then $\sigma_B^t(R,h^t,v_B^t) = \begin{cases} 1 & if \quad v_B^t \geq v_B^\ddagger = \alpha(R)\frac{k_B}{1-p\alpha(R)} \\ 0 & if \quad v_B^t < v_B^\ddagger = \alpha(R)\frac{k_B}{1-p\alpha(R)} \end{cases}$

2) If $h^t \in cheat^t$ then $s_B^t(h^t,M,v_B^t) = \begin{cases} 1 & if \quad v_B^t \geq \underline{v}_B = \frac{k_B}{1-p} \\ 0 & if \quad v_B^t < \underline{v}_B = \frac{k_B}{1-p} \end{cases}$, for all M.

If $h^t \in honest^t$ then $s_B^t(h^t,R,v_B^t) = \begin{cases} 1 & if \quad v_B^t \geq \widehat{v_B} \\ 0 & if \quad v_B^t < \widehat{v_B} \end{cases}$, and $s_B^t(h^t,S,v_B^t) = \begin{cases} 1 & if \quad v_B^t \geq \underline{v}_B = \frac{k_B}{1-p} \\ 0 & if \quad v_B^t < \underline{v}_B = \frac{k_B}{1-p} \end{cases}$.

3) If $h^t \in cheat^t$ then $s_A^t(h^t,M,v_A^t) = \begin{cases} 1 & if \quad v_A^t \geq \underline{v}_A = \beta\frac{k_A}{1-\beta+\beta p} \\ 0 & if \qquad\quad v_A^t < \underline{v}_A \end{cases}$, for all M.

If $h^t \in honest^t$ then $s_A^t(h^t, R, v_A^t) = \begin{cases} 1 & if \quad v_A^t \geq \widehat{v_A} = \beta(R)\frac{k_A}{1-\beta(R)+\beta(R)p} \\ 0 & if \quad v_A^t < \widehat{v_A} = \beta(R)\frac{k_A}{1-\beta(R)+\beta(R)p} \end{cases}$ and if $h^t \in Honest^t$ then $s_A^t(h^t, S, v_A^t) = 1$ for all v_A^t.

4) If $h \in Cheat^t$ then $s_E^t(h^t) = 0$.

If $h \in honest^t$ then $s_E^t(h^t) = \begin{cases} 0 & if \quad M = R, \text{ and } B \text{ surrenders} \\ 1 & \text{otherwise} \end{cases}$.

Where the terms $v_B^\ddagger$, $\underline{v}_B$, $\widehat{v_B}$, $\underline{v}_A$, $\widehat{v_A}$, β, $\beta(R)$, We_h, and $We_c$are defined below.

If actors play ACR then we can characterize observable behavior as follows: Let β represent the probability that B resists given a dishonest reputation: $\beta = 1 - F_B(\frac{k_B}{1-p})$. The probability that A attacks given a dishonest reputation is $\alpha = \Pr(v_A \geq \underline{v}_A) = 1 - F_A(\underline{v}_A)$, where $\underline{v}_A = \frac{\beta k_A}{1-\beta+p\beta}$. If B has an honest reputation and sends message R then the probability of resistance, $\beta(R) = \Pr(v_B^t \geq \widehat{v_B} | v_B^t \geq v_B^\ddagger) = \begin{cases} 1 & if \quad \widehat{v_B} \leq v_B^\ddagger \\ \frac{1-F_B(\widehat{v_B})}{1-F_B(v_B^\ddagger)} & if \quad \widehat{v_B} > v_B^\ddagger \end{cases}$, if B sends message S then the probability of resistance is $\beta(S) = \Pr(v_B^t \geq \underline{v}_B | v_B^t < v_B^\ddagger) = 0$. Given an honest reputation, the probability A attacks given message R is $\alpha(R) = 1 - F_A(\frac{\beta(R)k_A}{1-\beta(R)+p\beta(R)})$, and the probability of attack given message S is $\alpha(S) = 1$. The probability A attacks given a dishonest reputation is $\alpha = 1 - F_A(\frac{\beta k_A}{1-\beta+p\beta})$.

With a reputation for honesty, $h^t \in honesty^t$, the ex ante value for a citizen in B of playing a single period under the ACR strategy is $U_{e_B}(crisis | h^t \in honest^t) =$

$0F_B(v_B^\ddagger) + \int_{v_B^\ddagger}^{\widehat{v_B}} \left((1-\alpha(R))v_B + \alpha(R)\left(-\varepsilon\right)\right) f_B(v_B) dv_B +$

$\int_{\widehat{v_B}}^{\infty} \left(\alpha(R)((1-p)v_B - k) + (1-\alpha(R))v_B\right) f_B(v_B) dv_B$ if $\widehat{v_B} > v_B^\ddagger$ and $0F_B(v_B^\ddagger) +$

$\int_{\widehat{v_B}}^{\infty} \left(\alpha(R)((1-p)v_B - k) + (1-\alpha(R))v_B\right) f_B(v_B) dv_B$ if $\widehat{v_B} \leq v_B^\ddagger$. Without an honest reputation, $h^t \in cheat^t$, the ex ante value for a citizen in B of playing a single period under the ACR strategy is $U_{e_B}(crisis | h^t \in cheat^t) = E[U_B(crisis)] - \varepsilon$, where $E[U_B(crisis)]$ is the expected value of a crisis in the base case without informative signaling.

There are two cases of equilibria. Fully informative messages ($v_B^{\ddagger} \geq \widehat{v_B}$ and hence $\beta(R) = 1$) and partially informative equilibria ($v_B^{\ddagger} < \widehat{v_B}$ and hence $\beta(R) = \frac{1-F_B(\widehat{v_B})}{1-F_B(v_B^{\ddagger})}$). Again, the focus is on the former, not the latter.

Proposition: The Agent-Contingent Reputation Strategy is a perfect Bayesian Equilibrium, where with an honest reputation messages are fully informative of B's intention to resist ($\beta(R) = 1$, and $\beta(S) = 0$), if $\varepsilon \leq \delta(U_{e_B}(crisis|h^t \in honest^t) - U_{e_B}(crisis|h^t \in cheat^t))$ and $\delta \geq k_B \frac{1-\alpha(R)}{k_B(1-\alpha(R))+\Psi(1-p\alpha(R))}$.

The proof largely replicates the proof above for the CCR. The only decisions that require additional consideration are the citizens decision to remove dishonest leaders and the conditions under which leaders follow through on their declared intentions. Since both these decision are analyzed in the body of the text we omit a proof here.