Honest Threats

Matthew Draper

May 19, 2020

Contents

- Guisinger, Alexandra and Alastair Smith. 2002. "Honest Threats: The Interaction of Reputation and Political Institutions in International Crises". *The Journal of Conflict Resolution*, 46(2), 175-200.

Background

What is credibility? Guisinger and Smith advance the idea that credibility could attach not to states but to the individuals selected to represent them. "The benefits of deterring war through the use of diplomatic statements created its own value for a reputation for honesty" (175). The authors note that in crisis bargaining, both states have an incentive to arrive at a separating equilibrium where war only occurs when vital interests are threatened.

Background

They define reputation as a past record of diplomatic honesty (177). They seek to formalize the costs of having been "caught lying" in past diplomatic interaction. The authors believe that credibility is important because it allows states to avoid "inefficient" wars, which the authors define as wars that "would not be undertaken if the aggressor knew for certain that the target would resist" (178).

Background

They examine two models. In the first, called the "country-contingent reputation" model, states hold other states accountable for false diplomatic statements. In the second, called the "agent-contingent" reputation model, reputation attaches to leaders rather than to states.

The key distinguishing feature of these models is that they do not depend on 'resolve'. In addition, the authors argue that there is no interdependence between crises (though the tracking of reputation over time suggests otherwise). They are able to use these models to show that diplomatic communications have greater credibility and effectiveness under a broad range of conditions when leaders are domestically accountable (compare Fearon 1994).

Distinction from prior models

Fearon (1997) suggests that domestic audiences limit the scope of credible threats that a leader can make, because domestic audiences criticize leaders more for backing down after escalating a crisis than for not escalating in the first place. He suggests that this phenomenon will be more pronounced in democracies, and that as a result leaders of democratic states will have less room for bluffing, meaning that they will be better able to signal commitment.

Distinction from prior models

Guisinger and Smith point out that if we accept Fearon's argument, ex-post punishment of the leader is irrational because the costs of war would be worse, so the public shouldn't want to drive the state into war just to punish a bluffing leader.

They propose that the domestic audience punishes the leader not for bluffing (per se) but rather for "destroying the country's honest record" and thus jeopardizing the future benefits of credible crisis communication (179). They cite historical evidence on this point.[1] They also distinguish their models from work by Sartori (1998), which focused primarily on national-level reputation and did not endogenize credibility or model the constraints of domestic institutions.

[1] "Outside of political science, honesty is a common theme in French, British, and American manuals of diplomacy in every era (Nicolson [1939] 1964; Bailey 1968; Berridge 1995; de Callieres [1716] 1919; Cambon 1931)" (179).

Model 0: Crisis Interaction

Assume that a crisis exists between two states, A and B. B is satisfied with the status quo (it is in fact B's ideal position), while A is dissatisfied. If the status quo prevails, the payoffs are $(0, \nu_B)$. Conflict is a costly lottery where A wins with probability P. If A wins, the status quo payoffs are altered to $(\nu_A, 0)$. If B wins (probability $1 - p$), the states obtain the status quo payoffs minus the cost of conflict, k_A and k_B respectively.

Model 0: Crisis Interaction

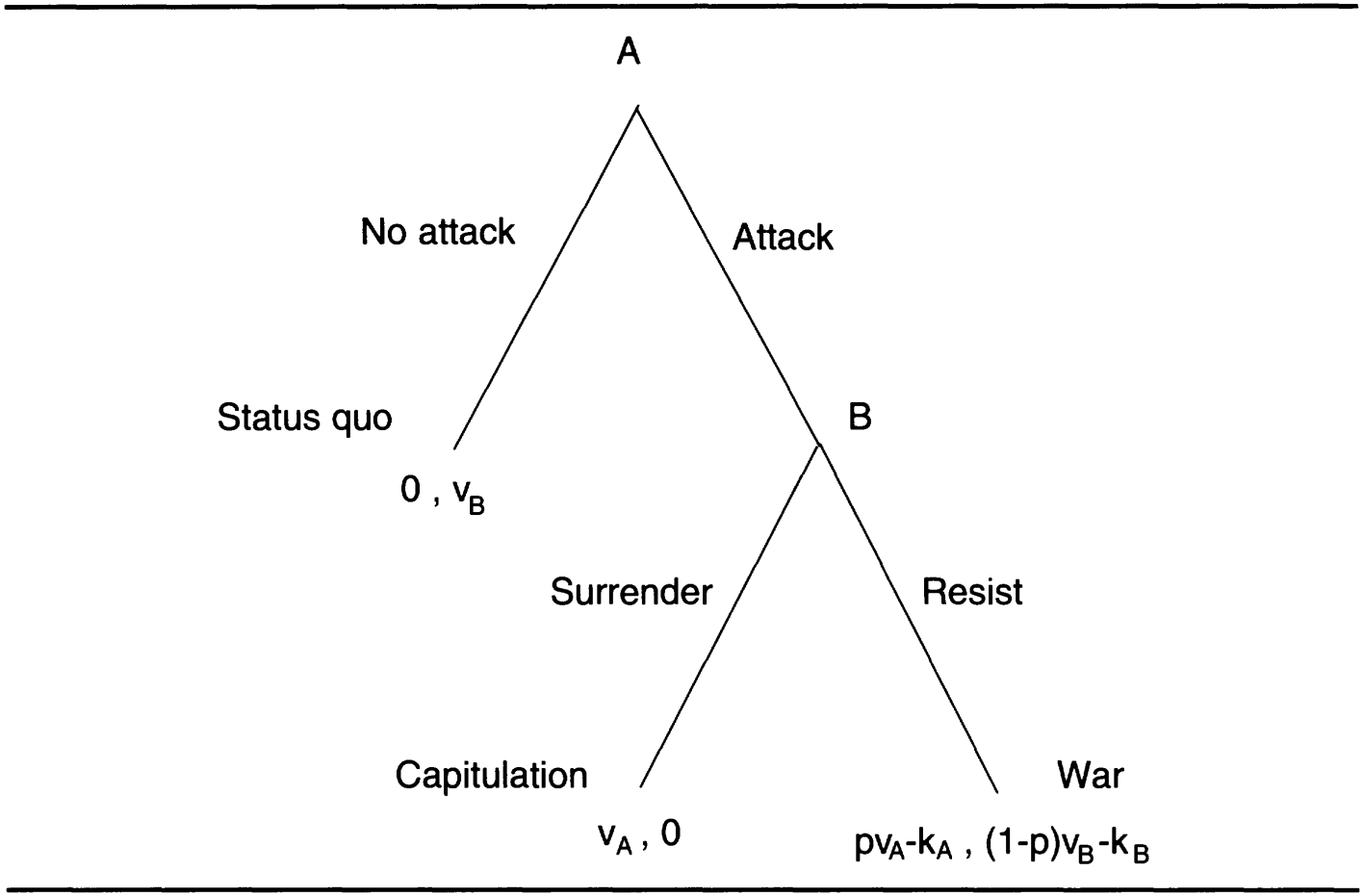

Figure 1: The Crisis Game

Model 0: Crisis Interaction

Crucially, to determine whether B will acquiesce or resist, A would need to know the value of ν_B. While for some issues A will attack regardless, Guisinger and Smith identify a class of wars that would not be undertaken if A could be sure about the value of ν_B. They say that in these latter cases, "war occurs needlessly" (182). We know that B's payoff from war will be:

$$E[U_B(resist|\nu_B)] = (1-p)\nu_B - k_B \tag{1}$$

So B will resist when the following inequality is true:

$$(1-p)\nu_B - k_B \geq 0 \tag{2}$$

$$\nu_B \geq \frac{k_B}{1-p} \tag{3}$$

Model 0: Crisis Interaction

Recall that A deos not know B's valuation of ν_B. A will therefore try to estimate its probability. Say that ν_B is distributed according to $F_B(x)$ (that is, $Pr(\nu_B \leq x)$ is $F_B(x)$). Call the ex-ante probability of B resisting β, which is ($Pr(\nu_B \geq \underline{\nu}_B)$, distributed uniformly according to F_B. Using this estimated probability, A can calculate the expected value of attacking, and A will attack if the following inequality is true:

$$\nu_A \geq \underline{\nu}_A = (1 - F_B(\frac{k_B}{1-p}))\frac{k_A}{p + (1-p)F_B(\frac{k_B}{1-p})} = \frac{\beta k_A}{1 - \beta + p\beta} \tag{4}$$

Model 0: Crisis Interaction

1. A's evaluation of the issue under dispute, $v_A > 0$, is randomly drawn from the distribution $F_A(v)$. A learns this value, but the members of country B do not, knowing only the distribution $F_A(v)$ from which it was drawn. The members of country B, the leader and electorate, simultaneously learn their valuation of the issue under dispute, $v_B > 0$. Again, country A knows only the distribution $F_B(v)$ from which v_B is drawn.
2. Leader B announces either the message R (an intention to resist) or S (no intention to resist).
3. Having observed the message R or S, country A decides whether to attack.[14]
4. If A attacks, then leader B decides whether to resist.
5. In the agent-contingent reputation (ACR) model, the citizens observe the outcome of the crisis and decide whether to retain their incumbent leader or replace him or her at a cost of ε.

Model 0: Crisis Interaction

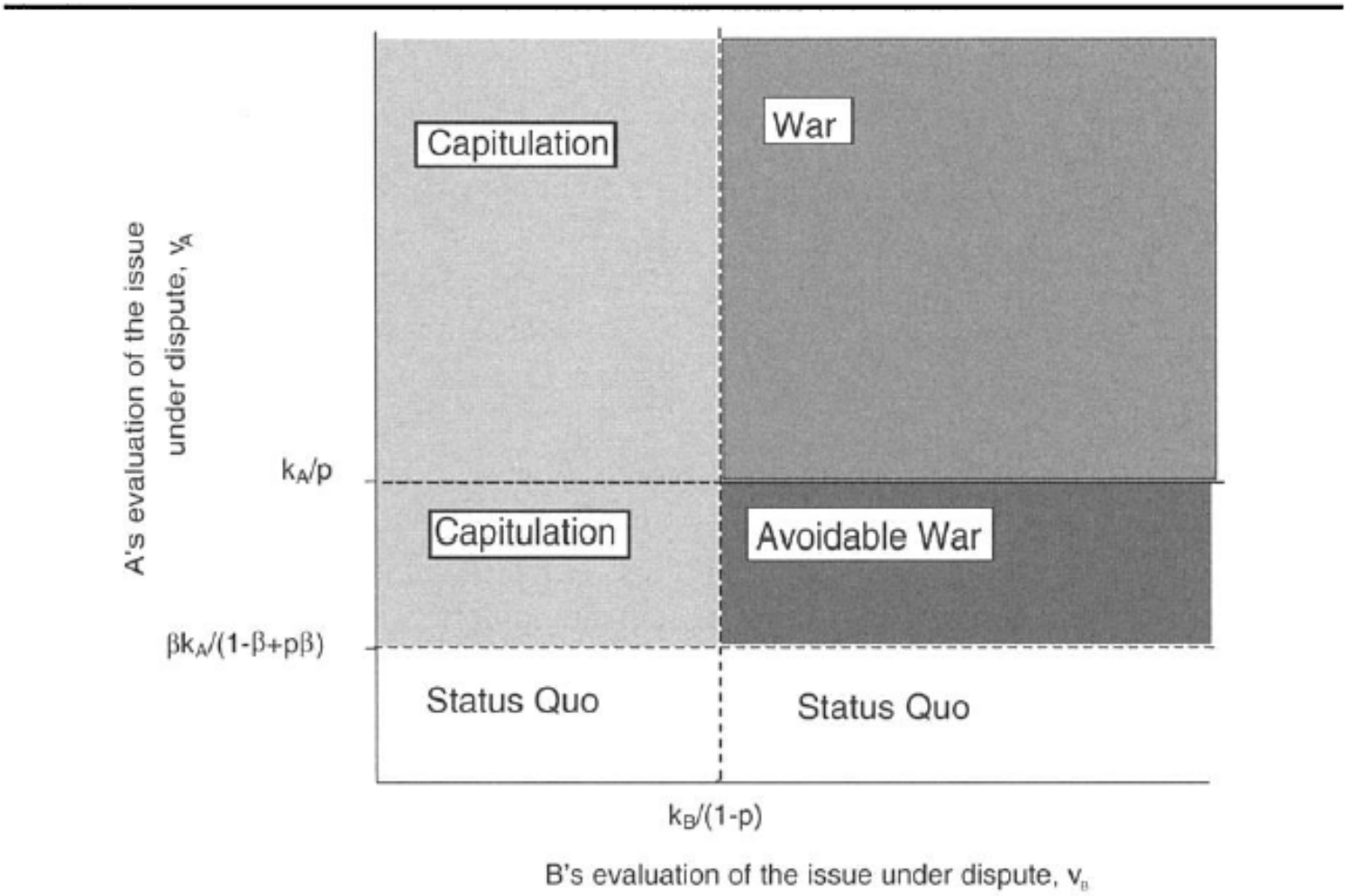

Figure 2: The Outcome in the Crisis Game Given the Type of Each Player

NOTE: B resists only if v_B is greater than $k_B/(1-p)$. A attacks, provided that v_A is greater than $\beta k_A/(1-\beta+p\beta)$. However, if A were certain B would resist, then A would only attack if v_A were greater than k_A/p.

Model 0: Crisis Interaction

Although A's and B's actions are ex-ante optimal, they are ex-post inefficient in the sense that sometimes A initiates an attack that leads to a war that A would prefer not to fight. Figure 2 depicts the strategies given the parameters $p = .5$ and $k_A = k_B = 0.2$.

If A knew that B would resist, A would only attack where $\nu_A \geq \frac{\beta k_A}{1-\beta+p\beta}$, but when A is uncertain, she will attack where $\nu_A \geq \frac{k_A}{p}$. In the numerical example used by the authors, A types above .17 attack when unsure about B's resolve, but only A types above .4 attack when A is sure that B is resolved.

Because B has no incentive to remain honest, she will signal resolve in all cases, effectively deterring all threats up to A's threshold of $\nu_A \geq \frac{\beta k_A}{1-\beta+p\beta}$. In this model, diplomacy is fruitless.

Model 0: Crisis Interaction

Recall that the authors define reputation as a past record of honesty in threats to challenge and resist (184). They assume that this reputation is common knowledge, and test two models: one where reputation is held at the level of the state (country-contingent reputation, or CCR), and another at the level of the leader (agent-contingent reputation, or ACR).

They assume that the values of the issue under dispute in each crisis (ν_A and ν_B) are re-drawn at the start of each period from the uniform distributions F_A and F_B. This endogenizes reputation, and states have no "underlying traits" (185). In addition, they introduce a round of pre-crisis communication in which B can announce its intention to resist (R) or surrender (S). A then decides whether to attack, and B must decide whether to resist or surrender (note that this need not track the message delivered in the pre-crisis communication).

Model 0: Crisis Interaction

Given the structure of the game, the only "history" of importance is B's reputation for honesty. Call the history of play h^t. The authors equate following through on threats to resist with an honest reputation ($h^t \in Honest^t$), and having made threats that were not followed through with a dishonest reputation ($h^t \in Cheat^t$).

Model 1: Country-Contingent Reputation

In the country-contingent reputation model, if B has an honest reputation A will believe B if B claims she is prepared to fight and hence conditions her action on B's diplomatic statements.

Let $\alpha(R)$ represent the probability that A attacks given that B has stated that she will resist (message R). Because this is a credible threat, A will only attack if she prefers the war outcome to the status quo (i.e. $p\nu_A - k_A \geq 0$), so $\alpha(R) = Pr(\nu_A \geq \frac{k_a}{p})$.

If B signals surrender (message S), A will always attack ($\alpha(S) = 1$). If B loses a reputation for honesty, the game collapses into the crisis interaction model (Model 0 above). Any decision by the electorate to remove the leader will be irrelevant.

Model 1: Country-Contingent Reputation

Consider an infinitely repeated version of this model. To calculate B's continuation value, call $\nu_B^{\ddagger}$ the value of ν_B at which B will resist rather than surrender.

In a given crisis, B will concede defeat if $\nu_B < \nu_B^{\ddagger}$, which we can express as a draw from the distribution $F_B(\nu_B^{\ddagger})$. In this case, A will attack, B will surrender and B's payoff will be 0, but B will preserve a reputation for honesty.

Model 1: Country-Contingent Reputation

With probability $1 - F_B(\nu_B^{\ddagger})$, B will value the issue enough to resist if challenged. In this case, B's payoff will be:

$$\alpha(R)((1-p)\nu_B - k_B) + (1 - \alpha(R))\nu_B \tag{5}$$

where the first term corresponds to the probability of A attacking multiplied by the expected payoff of conflict, and the second term is the value of the status quo multiplied by the probability that A will not attack.

Model 1: Country-Contingent Reputation

A will thus only attack if she prefers war to the status quo, or $\alpha(R) = Pr(\nu_A \geq \frac{k_A}{p})$. We can now state B's continuation value for playing H (not bluffing) on the uniform distribution as:

$$W_h = \frac{1}{1-\delta}(1-\nu_B^{\ddagger})(\frac{1+\nu_B^{\ddagger}}{2} + \alpha(R)(-p\frac{1+\nu_B^{\ddagger}}{2} - k_B)) \tag{6}$$

Model 1: Country-Contingent Reputation

Once a threat has been made, we must model the loss of reputation as a cost of not carrying it out. If a previously honest B is attacked following a declaration of intent to resist, its payoff for resisting (R) is $(1-p)\nu_B - k_B + \delta W_h$, where W_h is the continuation value of playing the infinitely repeated game with an honest reputation.

Model 1: Country-Contingent Reputation

If the same B were to choose to surrender (S), its payoff would be $0 + \delta W_c$. We can therefore say that types valuing the current issue more than $\hat{\nu}_B = \frac{k_B}{1-p} - \frac{\delta(W_h - W_c)}{1-p}$ will carry out their threats to resist to maintain an honest reputation. As long as $\nu_B^{\ddagger} \geq \hat{\nu}_B$, this is a perfect Bayesian equilibrium, but if not, then B cannot credibly commit to follow through on its threats. To ensure that $\nu_B^{\ddagger} \geq \hat{\nu}_B$, states must be sufficiently patient to value long-term reputation above the gains resulting from short-term defection.

Model 2: Agent-Contingent Reputation

Assume that reputation attaches to leaders rather than states, and that leaders enter office with a reputation for honesty (the authors say "a clean slate", but this is what it amounts to).

Assume that a state's electorate must pay some cost ϵ to remove and replace a leader, and that leaders obtain a benefit of holding office Ψ, independent of the outcome of conflcit. As in the CCR model, if B does not have a reputation for honesty play collapses into Model 0.

Model 2: Agent-Contingent Reputation

However, given an honest reputation $h^t \in Honest^t$, B will send the message R if and only if the issue under dispute is sufficiently valuable: $\nu_B \geq \nu_B^{\ddagger} = \alpha(R)\frac{k_B}{1-p\alpha(R)}$. A will believe threats to resist, and will only attack when it prefers war to the status quo: $\nu_A \geq \frac{k_A}{p}$, implying that $\alpha(R) = 1 - F_A(\frac{k_A}{p})$.

If B threatens to resist but is still attacked, then B's leader will choose to resist as long as $\nu_B \geq \hat{\nu}_B = \frac{k_B}{1-p} - \frac{\delta\Psi}{(1-\delta)(1-p)}$. Messages sent during pre-crisis communication will be credible as long as $\hat{\nu}_B \geq \nu_B^{\ddagger}$.

Model 2: Agent-Contingent Reputation

This model tracks the CCR model in many ways, except that a loss of reputation by B means the loss of ability to communicate in future crises, collapsing the payoff stream to the Model 0 case. But the electorate can restore the state's reputation by replacing the leader, and the threat of this punishment off the path of play will change the leader's incentives during crisis bargaining because she fears losing her payoff of Ψ.

Model 2: Agent-Contingent Reputation

As long as leaders care sufficiently about their payoffs from leadership, and the benefits of clear communication in future crises outweigh the cost to the electorate of removing a leader (ϵ), then we can reach perfect Bayesian equilibrium as long as:

$$\hat{\nu_B} \geq \nu_B^{\ddagger} \tag{7}$$

which implies:

$$\delta \geq k_B \frac{1-\alpha(R)}{k_B(1-\alpha(R)) + \Psi(1-p\alpha(R))} \tag{8}$$

$$\epsilon \leq \delta(U_{eB}(crisis|h^t \in Honest^t) - U_{eB}(crisis|h^t \in Cheat^t)) \tag{9}$$

where U_{eB} gives the utility to the electorate of playing a single crisis with an honest or dishonest leader, respectively.

Model 2: Agent-Contingent Reputation

Because the electorate threatens to remove leaders caught lying, the leader's incentives to honor her commitments is increased. As the value of office (Ψ) rises, more types can credibly commit to resist. Once $\nu_B^{\ddagger} \geq \frac{k_B}{1-p} - \frac{\delta\Psi}{(1-\delta)(1-p)}$, then all types who send message R subsequently resist. With a high value of Ψ, even relatively impatient leaders can credibly commit to action.

Model 2: Agent-Contingent Reputation

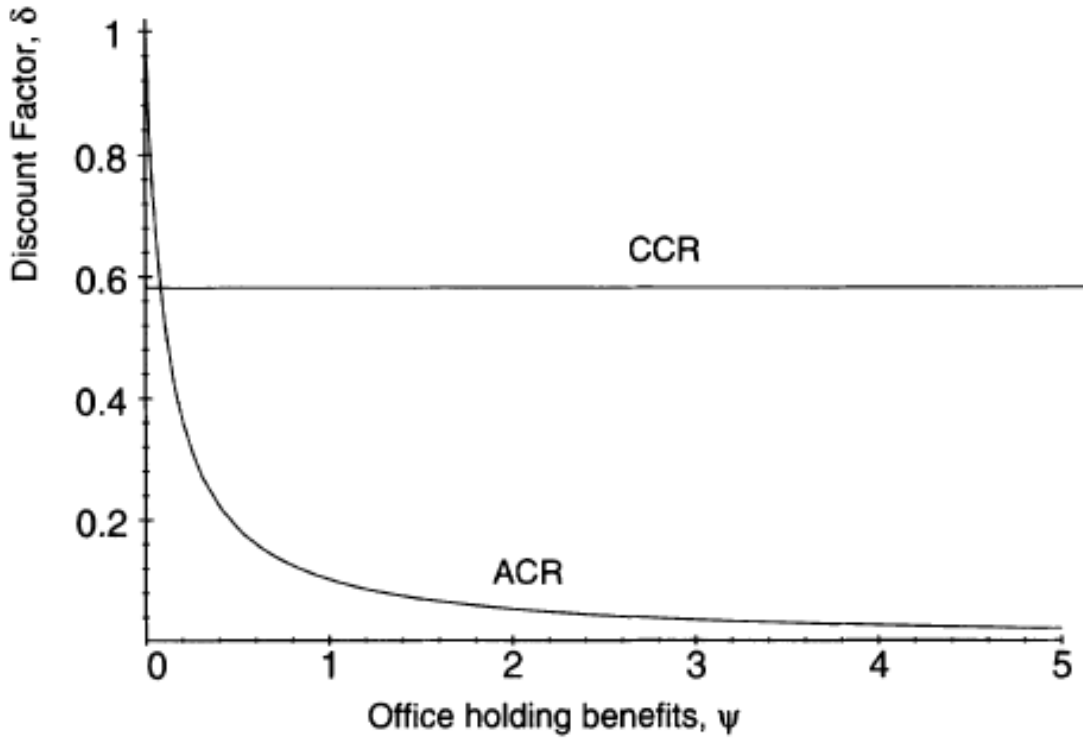

Figure 3: The Minimum Discount Factor to Ensure Fully Credible Foreign Policy Statements under the Country-Contingent Reputation and the Agent-Contingent Reputation Strategies

NOTE: Parameters: $k_A = 0.2$, $k_B = 0.2$, $p = .5$ and v_A and v_B uniformly distributed over the unit interval.

Discussion

The authors interpret both models as showing that diplomacy should be more effective than realists had supposed. In contrast with models where reputation is based on resolve, in these models states with a reputation for honesty will seek to protect it by avoiding commitments when they place little value on the issue under dispute (193). They claim that these models "resolve the paradox in Fearon's audience costs model as to why the domestic audience would ex-post punish a leader caught bluffing (194).

Discussion

They derive four major testable implications:

1. Provided that leaders care about office holding, the foreign policy statements cally accountable leaders are credible under a wider range of conditions than the tions of their autocratic counterparts.
2. Domestically accountable leaders are more likely to carry out any threats they make and hence are more careful to avoid making threats they are not prepared to carry out.

Discussion

1. The arena in which diplomatic communications take place depends on domestic accountability. Domestically accountable leaders use the public forum of press conferences, international summits, and direct public addresses to signal commitment policy. In contrast, public communiques by autocrats are unnecessary.
2. In general, the domestic accountability of democratic leaders means that they can more reliably signal their intentions, resulting in democracies being attacked less frequently, participating in fewer unnecessary wars, and benefiting from shorter negotiated settlements.

Discussion

Guisinger and Smith are certain that lies undermine reputation. I am not so sure. It seems to me that it is a reputation for lying that undermines reputation, not lies themselves. If Guisinger and Smith are right, then Machiavelli is wrong to urge leaders to cultivate a reputation for virtue but to take the ruthless actions necessary to benefit the state. While the model itself can accommodate this interpretation, the authors' substantive interpretation of it on the basis of diplomats' memoirs is questionable—leaders could have been lying their faces off even as they insisted on the importance of an honest reputation.

Discussion

A close examination of eighteenth century diplomacy in Europe suggests that this is in fact what was going on. Talleyrand was not known for his punctilious honesty, but he could keep commitments when it was in the national interest and his own simultaneously. Metternich and Castlereagh similarly dissembled when it was in their interest to do so, but could be relied on to pursue the national interest, which meant cultivating enough of a reputation for virtue to retain credibility.

However, to equate this desire to retain credibility with a penchant for truth-telling seems naive. We should not trust the memoirs of these diplomats when they assert the value of honesty. They have strong reasons to assert the value of honesty, and to cultivate reputations for it. I do not see why we should assume that this will lead them to behave honestly.