RATIONAL DETERRENCE IN AN IMPERFECT WORLD

By BARRY NALEBUFF*

I. Introduction

THIS paper applies recent economic research on games with incomplete information to the calculus of deterrence.[1] We do not attempt to provide a complete theory of rational deterrence. The focus is on the role of reputation and signaling in establishing deterrence. We show how strategic players signal a reputation for strength from the perspective of an internally consistent equilibrium model.

The theory of rational deterrence is based on the application of cost-benefit analysis to conflict initiation. The costs of conflict are compared with the benefits of cooperation; if all parties prefer cooperation, then the status quo will prevail and a potential conflict is avoided. Even at that level of generality, the theory is incomplete. A critical missing ingredient is the role of expectations.

A country's valuation of initiating a conflict depends on its belief about an adversary's intentions. This belief in turn depends on the perception of the adversary's own cost-benefit calculations. The fundamental problem is that all of these calculations and expectations are based on imperfect and incomplete information. The costs and benefits calculations combine subjective and objective elements. Parties in conflict are likely to view the situation from very different perspectives. How does rational deterrence apply when there is no objective set of calculations on which all parties can agree? In this paper, we show how to proceed with the necessary cost-benefit calculations in the presence of incomplete or imperfect information. The method is based on looking for a set of self-confirming beliefs. Consequently, there is no longer any guarantee that the calculations will provide a unique answer. In spite of this indeterminacy, we believe that the result is a much richer theory of rational deterrence.

The fundamental conceptual difficulty is the necessity of forming be-

* This paper has benefited from the generous comments of Michael Intriligator, Robert Jervis, John Vickers, and my SOM Thursday lunch colleagues. I thank the Pew Charitable Trust for financial support through their grant to the Princeton Center of International Studies.

[1] In truth, the calculus of deterrence might better have been called algebra.

liefs in the event that something which should never happen happens. For example, imagine that the cost-benefit calculation indicates that deterrence should work; then what do we believe about an adversary if deterrence fails? It would seem that a failure of deterrence indicates a failure of the model, and hence using the model to form expectations would seem to be a futile exercise. This is false. One must imagine the counterfactual possibility in order to perform the cost-benefit comparison that shows why deterrence should succeed.

The predicted probabilities derived from a rational deterrence model are endogenous and thus fundamentally different from the probability associated with the meteorologist's forecast of rain. If the weatherperson predicts a 100 percent chance of rain and it fails to rain, this should serve as conclusive evidence that the forecaster's model is flawed. But if a model predicts that deterrence should be 100 percent successful, that prediction can only be made by considering the costs and benefits associated with the event that deterrence fails. Our procedure for calculating the endogenous probabilities considers the "impossible" event in order to demonstrate that it will never happen.

This paper has two objectives. One is to defang the reputation paradox of Jervis. The second is to explain and motivate recent refinements in equilibrium theory. Each of the theoretical results is illustrated with a simple arithmetic example. The resulting equilibria offer different perspectives on how to interpret potentially misleading reputations and suggest several resolutions to the reputation paradox.

Section II begins with a general discussion of reputation. The possibility that reputations can be strategically manipulated leads to a paradox. This paradox is described in Section III, using the *Mayagüez* rescue as a case study. One resolution to the reputation paradox is presented in Section IV, where we present the idea of a sequential equilibrium. Increasingly sophisticated methods of interpretation lead to the refinements of Section V. Section VI offers a brief conclusion.

II. Reputation

A critical part of any rational deterrence calculation is to input the relevant payoffs. What should we believe about others and what should they believe about us? If preferences and payoffs were all known, then payoffs would be objective and specification straightforward. But in the presence of uncertainty and with the possibility of misperception, the payoff inputs are no longer clear. There arises a need to signal our objectives and to interpret the signals of others. How we communicate our objectives

(or how we hide them) becomes an integral part of the rational deterrence calculations.

One primary component of communication is the use of reputation. No cost-benefit calculation can be done in isolation. Each incident becomes part of the history for the future. Consequently, the cost-benefit calculation that leads to certain behavior in the present reflects on how one will act in future cases. During the Berlin crisis in 1961 John F. Kennedy explained the importance of the U.S. reputation: "If we do not meet our commitments to Berlin, where will we later stand? If we are not true to our word there, all that we have achieved in collective security, which relies on these words, will mean nothing."[2]

The problem is that everyone has an incentive to talk tough. Making the correct inference is not straightforward. In a chapter entitled "Signals and Lies," Robert Jervis explains: "Whether the state is lying or honest it will try to issue those signals which it thinks will get its desired image accepted. . . . Both an honest man and a liar will answer affirmatively if asked whether they will tell the truth."[3]

Thus we judge actors by their actions rather than their words. Reputation, based on a long and consistent history of behavior, helps predict the future. The Mayflower furniture company along the Massachusetts Turnpike proudly advertises that it has gone 127 years without a sale. (Are they still waiting for their first customer?) This unconditional commitment to everyday low prices brings in a year-round steady stream of customers. A sale might temporarily raise profits, but it would be another 127 years before they could repeat such a clever advertisement. Next year, we expect the sign will read 128 years. The reputation becomes self-perpetuating as it becomes more valuable.[4]

Rarely do politicians or countries have the opportunity to establish such a long-standing reputation. Moreover, the application of the reputation is not nearly so much in their control. The reputation of the Mayflower furniture company is sufficient to deter one particular event, a sale. Countries are not so fortunate; reputations must be built to deter unforeseen future events.

A country's reputation is multidimensional, and yet there are rarely more than a few observations on which to base beliefs. The problem with integrating reputation into the cost-benefit calculations is that one must extrapolate out of the sample. How might President Kennedy's response

[2] Quoted in F. Ikle, *How Nations Negotiate* (New York: Harper and Row, 1964), 67.

[3] Jervis, *The Logic of Images in International Relations* (Princeton: Princeton University Press, 1970).

[4] Sadly, we must report that the Mayflower furniture store recently had its first sale, a going-out-of-business sale.

to Berlin forecast the U.S. response to a Soviet invasion of Europe? Did Kennedy's reputation extend to President Johnson? Could the U.S. invasion of Panama be predicted from the response in Grenada?

The problem is further complicated by the endogenous nature of the inferences. *The value of a reputation depends on how others interpret it.* This in turn affects the willingness to create the reputation in the first place. The strategic use of reputation cannot be evaluated in isolation. It must be part of an equilibrium model.

The advantage of a formal model is that it forces us to integrate the consistency requirements of equilibrium with the strategic use of signaling. Reputation is presented as a continuous variable. Consequently, the issue is no longer whether to believe but *how much* to believe based on what one sees. The equilibrium level of reputation defines the extent to which one may attempt to extrapolate the future from the past.

III. Signaling and Deterrence: A Paradox

To discuss the role of signaling and reputation as part of deterrence, we start with the work of Robert Jervis. His analysis has been at center stage beginning with *The Logic of Images in International Relations.* An interesting and seemingly counterintuitive theme from Jervis is that signaling may be counterproductive. A country may get stuck in a negative feedback cycle. He uses the *Mayagüez* rescue as a representative case study.

After the U.S. retreated from Vietnam, it was important for the Americans to reestablish a reputation for toughness. The rescue of the *Mayagüez* seemed to provide the perfect opportunity. On Monday morning, May 12, 1975, a U.S. vessel called the *Mayagüez* was captured by a Cambodian torpedo boat. The Cambodians accused the ship of spying.[5] President Ford first attempted to negotiate the release of ship and crew. But there was no diplomatic response from the Cambodians. Instead, they fired at the U.S. reconnaissance planes flying over the *Mayagüez.* On Wednesday evening President Ford sent in the marines. Although there were heavy marine casualties (partly due to poor intelligence about the ship's location), the captain and his thirty-eight crewmen were rescued. The *New York Times* reported as follows:

> [Both politicians and the public] saw the event as a re-assertion of American will after this country's disorderly retreat from Indochina. ... Ever since the American evacuation from Saigon, Administration officials had

[5] But after an interpreter was found, the crew did not appear to be in immediate danger. The Cambodian captors were apparently persuaded that the *Mayagüez* was not a spy boat.

> been saying frankly that America's international stature could be restored by a demonstration of strength.[6]

While the stated purpose was solely to rescue the *Mayagüez* crew, even the headlines read "Ford Sends a Signal." The *New York Times* article continued: "The Administration has been specific about one nation, North Korea, to which it would like to send a clear, strong signal since the defeat in Indochina. Now they say the signal has been sent: Don't make a move against South Korea without expecting American military intervention."

Was it rational to expect that rescuing the *Mayagüez* would restore the U.S. reputation for strength? If so, then the *Mayagüez* rescue would have been done even by a "weak" U.S. government, for this would be a small price to reestablish its reputation. But then observers should realize that intervention is no longer a signal of strength. Nothing is learned from the American behavior, since both a tough and a weak U.S. would pursue the same strategy. Jervis explains the difficulty in drawing inferences:

> This raises [a] problem with such inferences about resolve. Beliefs like these will undermine rather than support themselves: if the U.S. felt that fighting a small war for a country of little intrinsic value would lead others to conclude that it would display high resolve in a dangerous confrontation, then this action would not provide reliable evidence because the U.S. would fight the small war in order to create a favorable impression irrespective of whether or not it would run high risks in a nuclear exchange.[7]

If we take this argument to its logical conclusion, the U.S. allies (and adversaries) should pay no attention to the U.S. theatrics of acting tough. It has no signaling value. This brings us around full circle. If acting tough does not improve one's reputation, then what is the reason for a weak party to pretend to be tough? The negative feedback seems to dissipate any possible improvement in reputation.

Improvement is, of course, relative. While the reputation may not improve relative to the status quo, the status quo is no longer an option. Thus we care about the relative effect of intervening on reputation as compared with not intervening. A country that fails to act could suffer a massive loss in reputation. Avoiding the loss is what keeps the cost-benefit calculations positive.

The problem with this interpretation is that there is a lack of discipline in determining what one can and cannot believe. In order to determine

[6] *New York Times*, Week in Review, May 18, 1975.

[7] Jervis, "The Symbolic Nature of Nuclear Politics" (The Edmund James Lecture, Department of Political Science, University of Illinois at Urbana-Champaign, 1985).

when a party should intervene, we must simultaneously determine when it is best not to intervene. Neither calculation can be done in isolation.

The intuition behind the formalization of equilibrium is that expectations are in fact negatively equilibrating. The more desirable a certain action is, the less one infers about those who take the action. There is less of an effect on reputation and consequently there are fewer circumstances in which the action is taken. This helps restore the implications about those who act. In equilibrium, the effect on reputation is just sufficient to motivate those who act and no others.

Note that an essential part of this story is that reputations are continuous. There is the possibility of drawing greater or lesser inferences. It is this move away from black and white beliefs that allows us to find an equilibrium in spite of the negative feedback effect. To see the problem with black and white beliefs, we return to an analysis from Jervis. Read it with an eye toward calculating what one should believe, not what one should not. Jervis writes:

> An ironic possibility should be noted. A concern for reputation can lead states to act and draw inferences in a pattern opposite from the one that we—and most other analysts—imply. This is not to dispute the common starting point; states refuse to back down not because of the immediate and direct costs of doing so, but because of the belief that a retreat will be seen as an indication of general weakness and so lead others to expect retreats in the future. But the desire to counteract such undesired consequences may lead a state that has retreated on one issue to pay especially high costs to avoid defeat on the next one. Thus the United States was not only willing but anxious to use force to free the *Mayagüez* because it wanted to show others that its evacuation of Indochina didn't mean it would not defend its other interests—the very consequence which it had predicted would follow from a defeat in Vietnam and which had justified its participation in the war. If others understand this logic and expect states to behave in this way—to follow retreats with displays of firmness—then reputations for carrying out threats do not influence estimates of credibility because—to compound the paradox—reputations are so important that states must rebuild them when they are damaged. If you have been caught bluffing in poker, others are likely to call you in the next round in the belief that you bluff a lot or are they unlikely to do so because they think you know it is no longer safe to bluff? To the extent that the latter is the case, perceptions of credibility are influenced by the state's recent behavior, but in a way which produces equilibrating negative feedback rather than the positive feedback of domino dynamics.[8]

The poker story shows the seeming paradox of self-falsifying inferences. If one believes that a player who recently bluffed will now play

[8] Jervis, "Deterrence and Perception," in Steven Miller, ed., *Strategy and Nuclear Deterrence* (Princeton: Princeton University Press, 1984), 66–67.

safe in order to reestablish a reputation, then there is an incentive to continue bluffing. But if this is the belief, then others will continue to call and the player should play his cards straight. Neither inference can be an equilibrium. What is?

One option left out is the possibility of a randomized (or mixed-strategy) solution. The need for a mixed strategy is really a consequence of the artificial assumption that there are only two types of actors, weak and strong. To restore continuity, a reputation is based on a probabilistic belief about an actor's type—a reputation is the inferred probability that the actor is strong.

Imagine that a strong country always acts tough. The reputation effect from acting tough will depend on what a weak country is expected to do. If it was thought that a weak country would never attempt to bluff or act tough, then seeing a country act tough would indicate true strength. This would improve its reputation immensely, possibly enough even to motivate a weak country to act tough (which would not be an equilibrium). As bluffing becomes increasingly likely, the enhancement of a reputation following tough behavior is diminished. At some point, the probability of bluffing is sufficiently high (and the improvement in reputation is sufficiently small) that the cost of the weak country acting tough is exactly offset by the gain in reputation. This is the mixed-strategy equilibrium.[9]

Although the mixed-strategy approach provides a consistent solution, it may have a somewhat artificial flavor. This is solely a consequence of the highly stylized assumption that there are only two types of actors, weak and strong. The problem is both simpler and more realistic when we allow for a continuous range of incomplete information.

[9] Although we have found a consistent solution, is it of any use? Here Jervis has some doubts as to the ability of game theorists to say much more about the "black box" of mixed strategies:

> Economists have not been able to model the behavior of oligopolists nearly as deterministically as they have that of the wheat farmer facing a market he cannot influence. . . . [I]n many situations game theory prescribes a mixed or randomized solution. Of course, game theory yields great insights into how actors try to out-think and out-bluff each other, *but the competitive and variable sum nature of the situation means that scholars cannot produce deductions on the model*: "In situation X the actor always should (or will) do Y." We can show that the actors' calculations are consistent with deterrence, but this requires looking into the "black-box" of decision-making. (Jervis, "Rational Deterrence: Theory and Evidence," *World Politics* 41 [January 1989], 183–207)

In fact, it is possible to characterize the precise equilibrium proportion of mixing. Or to put it more poetically, there is a method to the madness of a random strategy; see Avinash Dixit and Barry Nalebuff, *Thinking Strategically: A Competitive Edge in Business, Politics, and Everyday Life* (New York: W. W. Norton, 1991). To verify whether or not actors follow their prescribed mixed-strategy rules does not require a large number of independent observations from a repeated game. Instead the theory is tested by evaluating how well it predicts cross-sectionally, looking across a variety of different conflict situations.

IV. Equilibrium

There is another way to tell the *Mayagüez* story. Instead of the artificial nature of just two types, weak and strong, imagine that the propensity to intervene comes in continuous varieties. To be more concrete, we consider a stylized model of conflict initiation.

There is an event that provides the U.S. with a payoff of x if it intervenes. The variable x should be thought to include all the observable elements of the cost and benefit calculations. In addition to the observable payoffs, there is an unobservable component, c. The variable c can be thought to represent all the psychic and other intangible costs and benefits associated with intervention. For ease of exposition, x is described as a benefit and c as a cost.[10] Thus, a low value of c indicates a greater propensity to intervene, as the unobserved costs are low. Conversely, a high value of c is evidence against intervention. Intermediate values of c indicate intermediate willingness to intervene.

The role for a reputation arises because countries do not know each other's c. Each side starts with some expectation about the distribution of the other's parameter. For analytic convenience, we take the initial beliefs to be uniformly distributed between zero and one.[11] These beliefs are updated based on observed behavior. Thus a country may act strategically in order to manipulate how others perceive its intervention cost.

Without any notion of reputation, the U.S. would intervene if $x > c$ and do nothing otherwise. However, because the U.S. cares what others think about its unobserved parameter, the effect on reputation enters into its cost-benefit calculations. Once again for convenience, let the value of a reputation equal $a\ [1 - \bar{c}]$. Here $\bar{c}$ is the average value others think the c is for the U.S. and a is a parameter that measures the importance of reputation. One interpretation of the reputation effect is that the unobserved costs and benefits are correlated over time and across different circumstances. Reputation captures the future value of changing others' perception of your unobserved costs.

Initially, $\bar{c} = 1/2$. When $a > 0$, the U.S. wants others to believe that its value of c is low since this will make them less likely to act against the U.S. interest. The parameters a and $\bar{c}$ provide a shorthand, or reduced form, representation for the value of a reputation. A larger value of a corresponds to a greater significance placed on reputation. A change in $\bar{c}$ corresponds to a revised belief about the country's cost of intervention.

[10] But there is nothing that prevents x and c from being negative, in which case the observed value would represent a net cost and the unobserved parameter would represent a net benefit.

[11] Here, the assumption that intervention costs are uniformly distributed is made solely for analytic convenience. The prior belief should be based on the history up to this point.

To model the outcome, we begin with a listing of the minimal requirements for an equilibrium. Denote the set of types who intervene by I and those who choose not to intervene by N. There is an expectation about the representative type of country that chooses to intervene; this is denoted by c_i. For a country that chooses not to intervene, the expectation of its cost is c_n.[12]

1. *Maximizing behavior.* (a) The payoff to any country that intervenes is higher than if it does not intervene; (b) the payoff to any country that does not intervene is higher than if it does intervene.

2. *Consistency of beliefs.* (a) If the sets I and N are both nonempty, then the expectation of the costs of those in each set should be based on the distribution of types in each set; (b) if one set is empty (say N) and the other contains everyone (say I), then expectations about those in I must equal the prior belief, but there is freedom to form an expectation about who might be in N should this zero-probability event arise.

Together these requirements are called a sequential equilibrium, a refinement of Nash equilibrium due to Kreps and Wilson.[13] It is important to emphasize that these two conditions are minimal requirements for an equilibrium. Maximizing behavior is an essential element of rationality. It imposes cost-benefit analysis as the basis for deciding whether or not to intervene. The consistency of beliefs condition is more subtle. The second part of this condition provides a degree of freedom that often permits a spectrum of equilibrium outcomes. Consequently, more than one equilibrium outcome may satisfy maximizing behavior and consistency of beliefs. Some are more appealing than others. To choose between equilibrium outcomes, one may look to impose a stronger test of rationality.

If, as part 2(a) supposes, we expect countries with one range of costs to intervene and those with some other range not to, then our conclusion about who did what should be based on the ranges corresponding to the observed action (or inaction). But what are we to believe about the set of noninterveners if the expectation is that everyone will intervene? In case 2(b) we are forced to have an expectation, c_n, over who might be in this null set. Sequential equilibrium is agnostic in this matter and allows the expectation to take any possible value. (For this expectation to be part of an equilibrium, the reputation effect must then motivate everyone to choose intervention as their maximizing behavior.) It is possible to form more sophisticated expectations about who might have taken an action

[12] Note that we do not require that $c_i \leq c_n$. It is possible that intervening hurts one's reputation. Such an example is presented below.

[13] David Kreps and Robert Wilson, "Sequential Equilibria," *Econometrica* 52 (July 1982), 863–95.

that never should have happened. In Section V, we provide three increasingly restrictive assumptions about how to form rational expectations in case 2(b). The three refinements of sequential equilibrium we consider are the successive elimination of dominated strategies, universal divinity, and perfect sequential equilibrium. Each is defined and illustrated using the reference example. We begin with the possibilities for a sequential equilibrium.

There is an event that gives the U.S. a chance to improve (or worsen) its reputation. Intervention is observed to be worth an amount x. To find a consistent set of beliefs, we undertake the cost-benefit analysis assuming there is some expectation about the unobserved cost for those who intervene and those who do not. These expectations lead to a reputation c_i for a country that intervenes and a reputation c_n for a country that maintains the status quo. With these reputations in mind, the U.S. chooses its optimizing behavior. The resulting behavior is an equilibrium if and only if the optimization confirms the initial expectations about the expected intervention cost for the U.S. when it intervenes and when it does not.

Were the U.S. to intervene when its true cost was c, its payoff would be

$$x - c + a[1 - c_i].$$

The U.S. earns an observed payoff of x, pays an unobserved amount c in intervention costs, and ends up with a reputation valued at $a[1 - c_i]$.

If the U.S. does not intervene when its true cost is c, its payoff is then

$$a[1 - c_n].$$

Note that this valuation is independent of x and c since x is not earned and c is not paid; instead, the U.S. ends up with a reputation valued at $a[1 - c_n]$.

Hence, it is better to intervene provided that

$$c < x + a[c_n - c_i].$$

We call this critical cost c^*,

$$c^* = x + a[c_n - c_i].$$

As seems intuitive, when intervention costs are low [$c \leq c^*$], the U.S. acts, whereas when intervention costs are high [$c > c^*$], the U.S. does not act.[14]

The calculations above provide a formula that reveals the critical value

[14] How we decide the action for the c^* type is irrelevant.

c^* for any imagined values of c_i and c_n. Consistency requires that we expect an intervener to have costs in the range [0, c^*] and that we base our belief c_i on the prior distribution of costs in that interval. Similarly, noninterveners have costs in the range (c^*, 1], and so the expected value of c_n is based on the prior distribution of costs in this upper range. It is important to note that this consistency condition falls under 2(a) only when $0 < c^* < 1$. The two other possibilities are discussed in depth below.[15]

Because the prior distribution is uniform, when intervention is observed, the expected value c_i taken over the range [0, c^*] must be

$$c_i = c^*/2.$$

If no intervention takes place, then the expected value for c_n is

$$c_n = [1 + c^*]/2.$$

We have two equations and two unknowns. Any solution is a sequential equilibrium. Taking the difference between the two equations shows that for any solution, $[c_n - c_i] = 1/2$.[16] We first solve for c^* and then c_i and c_n.[17]

$$c^* = x + a/2, c_i = (x + a/2)/2, c_n = (1 + x + a/2)/2.$$

To provide a numerical illustration, let $a = 1/2$, and consider an event with $x = 1/4$. Then $c^* = 1/2$, $c_i = 1/4$, and $c_n = 3/4$. There is a sequential equilibrium in which all the types with intervention cost below 1/2 act whereas those with costs above 1/2 do nothing. Moreover, ruling out for the moment solutions with zero-probability events, this equilibrium is unique.

As the parameter values change, the predicted outcome varies in a natural way. A rise in the value of reputation, a, results in greater participation. A higher reward, x, from intervening also encourages greater intervention in equilibrium.

This simple model illustrates the cost of maintaining a reputation. The only way to prove something about yourself is to take an action that would be too costly if you are not who you are trying to convince others

[15] If $c^* = 0$ then no one should intervene, whereas if $c^* = 1$, then everyone should intervene. Both endpoint cases predict that one action is never taken and hence falls under the weaker consistency condition 2(b).

[16] This simple relationship is an artifact of the uniform distribution of intervention costs.

[17] This type of well-behaved sequential equilibrium exists provided $0 < c^* = x + a/2 < 1$. Otherwise, the effect of reputation may be so large or the value of intervening so great that we expect everyone to act, $c^* \geq 1$. Alternatively, the value of reputation may be so small (or negative) or the price of intervention so costly that all types choose not to intervene and $c^* \leq 0$.

you are. But unlike the circular reasoning of Jervis, this can be done in a matter of degree. The unobserved cost of intervention separates the weak from the very strong—not perfectly but some part of the way. Seeing a country intervene does not reveal its exact cost, c; it just sends a signal that the intervention costs are below rather than above c^* and thus average c_i rather than c_n.

An implication of the model is that we see too much intervention. Here too much intervention means that intervention takes place for its effect on reputation and would otherwise not be desirable. An intervener's gain in reputation is a nonintervener's loss. Put together the two effects exactly cancel out, and we have a zero-sum transfer. This excess intervention would not occur if the country's unobserved parameter were known. A country would intervene only if $x > c$. There would be no reputation effect since if c were known, its perception could not be changed.

When costs can only be inferred rather than observed directly, then the type with $c = x$ strictly prefers to intervene (and by continuity so will types with $c > x$). The reason is that intervention has little or no direct cost and results in a positive gain in reputation for being perceived as a c_i rather than a c_n.

The reputation effect is a two-edged sword. As Jervis emphasizes, a country that does not act cannot presume its status quo reputation remains unchanged. That effect is captured in the cost-benefit analysis. A country that fails to intervene sees the value of its reputation fall, not necessarily from the status quo but from the value to which it would have risen had it intervened. In our example, the status quo value of a reputation is 1/2. But once the opportunity to intervene arises, those who act are viewed as having an expected cost $c_i = 1/4$ while those who do not intervene are represented by an average cost $c_n = 3/4$—there is no option to remain at $\bar{c} = 1/2$.

Self-fulfilling Equilibria

So far the calculations have been straightforward. Now we turn to the more subtle possibility of an equilibrium based on self-fulfilling expectations. In these cases everyone is expected to choose the same action, either intervention or nonintervention. No one dares to be different because the loss of reputation associated with taking the "wrong" action dominates the direct gain or loss.

The possibility of multiple equilibrium based on different expectation foundations is described in Jervis: "There is a great deal of room for false

consensus effects—i.e., if each person thinks everyone else holds a certain view, then that view becomes the operating reality. . . . The truth of the statement depends on whether people believe it."[18]

This is the problem we now confront: to what extent can expectations lead to an equilibrium? What are the limits to which beliefs are self-fulfilling? It is here that the power of equilibrium reasoning has its greatest effect. We are able to show exactly what type of beliefs are internally consistent. The scope for expectations to drive the model is remarkably limited.

The nature of these self-fulfilling equilibria relies on what might be called a "lemmings" effect: it must be that *everyone* is expected to do the same thing in equilibrium.[19] There are two candidate equilibria. In one, everyone intervenes because the loss of reputation from not doing so overwhelms other considerations. The more types that act, the more costly it becomes not to act. The expectations feed upon themselves until even the weakest type is forced to intervene in order not to be exposed. This is just the limiting case of our previous example.

The other possibility is that no one intervenes; the reason is that an intervener is thought to be weak—not strong—so the reputation loss from intervening outweighs any direct gain. This reversal of a reputation effect is quite different from what we have previously discussed. To formalize these possibilities it is appropriate to return to the lurking issue of what to believe when we see something that we thought was not possible.

One interpretation upon observing a zero-probability event is that the model is wrong. Since we have seen something that obviously should never have happened, we must have misunderstood the original situation. This negative perspective misses the point of cost-benefit analysis. It is simply impossible to predict that something should not happen without analyzing the payoffs if it did. The choice of action is based on a calculation of net gain. How can we choose an optimal decision knowing only one side of the equation?

Rather than discuss this issue in the abstract, the model provides a revealing backdrop. Imagine that $c^* = x + a/2 \geq 1$. In this case we predict reputation is so important that everyone should be willing to intervene. But how is this justified? It is clear what we should assume if we see an intervention; since we expected everyone to behave this way,

[18] Jervis (fn. 7).

[19] Otherwise, there are no zero-probability events. The earlier analysis applies and there is a unique solution.

it should have no effect on our prior beliefs. Yet this calculation was conditioned on the cost-benefit calculation, which by necessity must place a value on not intervening.

We must specify how others will interpret a failure to intervene. Since nonintervention is a zero-probability event, we cannot apply the standard technique to calculate posterior beliefs. The standard technique used to calculate posterior beliefs about who did what is Bayes' rule. The formula for the posterior probability that someone has cost c conditional upon observing nonintervention is

$$\text{Prob}[\text{cost} = c \mid \text{nonintervention}] = \text{Prob}[\text{cost} = c \ \& \ \text{nonintervention}] / \text{Prob}[\text{nonintervention}].$$

The problem is that nonintervention is supposed to be a zero-probability event so that both the numerator and the denominator are zero. Bayes' rule makes no prediction in this case. The question of what is reasonable to believe in this surprise event is an unresolved research question in game theory. The present model helps illustrate some of the possibilities. Here it is important to emphasize that there is more than one way to form rational beliefs and that reasonable people differ over which to choose.

One solution to this paradoxical situation is provided in the definition 2(b) of sequential equilibrium. Choose any beliefs over the range [0, 1] for the types who do not intervene. The only constraint is that these beliefs must then be consistent with the initial assumption that no one would in fact want to intervene. The way to check if this is possible is to assign "pessimistic" beliefs in the event of this zero-probability outcome. For example, if the country fails to intervene, we believe that its unobserved costs are 1, the worst possibility. Given this belief, when is the model internally consistent? The payoff to intervention is $x - c + a/2$, while the payoff to nonintervention is $a(1 - c_n) = 0$. All types prefer to intervene when

$$x - c + a/2 \geq 0, \ 0 \leq c \leq 1.$$

This is true if and only if $x + a/2 \geq 1$.

Thus our intuitive belief that all types will intervene when $x + a/2 \geq 1$ is confirmed. When the observed payoff for intervention is large (compared with the unobserved costs) and reputation is highly valued, then it is natural to suppose that there is an equilibrium in which all countries intervene. Except for the issue of forming expectations in a zero-probability event, there is nothing different or unusual about this self-fulfilling equilibrium. The outcome is simply the limiting case of our previous example for $c^* \geq 1$.

The power of sequential equilibrium is that it restricts the extent to which expectations can drive the outcome. Believing in something cannot always make it happen. In our numerical example, there is *no* sequential equilibrium that involves intervention by all parties. The reason is that with $x = 1/4$ and $a = 1/2, x + a/2 < 1$; a country with intervention cost c close to 1 would prefer to have it presumed that $c = 1$ than to lose nearly $3/4$ in order to improve its reputation to $c_i = 1/2$.

The Reversed Equilibrium

We now turn to the paradoxical case. It is often possible to find a second set of beliefs consistent with the requirements of a sequential equilibrium. For certain parameter values there exists a sequential equilibrium in which no intervention occurs. This solution is fundamentally different in character from our previous examples. The out-of-equilibrium beliefs are reversed. We assume that a country that intervenes is *weak*, not strong. Restraint is the sign of strength. The interpretation is that a country that intervenes is so much trying to prove itself that it reveals its true weakness; countries with low unobserved costs are sufficiently confident of their capabilities that they are willing to forgo this intervention. Of course, this interpretation is particularly convenient for a weak country, which by doing nothing can pretend that it is so strong that it does not have to intervene.

Is it possible that these beliefs are consistent? To support an equilibrium where intervention is supposed to be a zero-probability event, we assign posterior beliefs that an intervener has $c_i = 1$. The unobserved costs assigned to an intervener are as large as possible. The payoff to a country if it intervenes is then

$$x - c + a(1 - c_i) = x - c.$$

Expectations following the case of nonintervention are more straightforward. The presumption is that no one should intervene. Hence, observing that a country does not intervene should have no effect on its estimated intervention cost. A country that does not intervene has its reputation preserved at the prior level, $c_n = 1/2$. The payoff for this strategy is then $a[1 - c_n] = a/2$. Not intervening is the best strategy for a country with unobserved cost c when

$$a/2 \geq x - c.$$

Since the presumption is that no one intervenes, this inequality must be true even for the country with the greatest propensity to intervene, the type with $c = 0$. This implies

$$x - a/2 \leq 0.$$

Again, consider our numerical example with $a = 1/2$ and $x = 1/4$. In this case, there is a sequential equilibrium in which no type intervenes. The reason is that the loss of reputation costs 1/4, which is just enough to offset the value of intervening for all types.[20]

The advantage of the formal modeling is to put a consistency constraint on what might be thought of as an equilibrium. The examples illustrate that sequential equilibrium does not go very far in that direction. The solution concept seems to allow opposite extremes just by changing the interpretation of a zero-probability event.

V. Refinements

Sequential equilibrium is just the first step on the road to forming beliefs out of equilibrium. It is a minimal set of requirements. Consequently, there may be a multiplicity of equilibria. In this section, we introduce three refinements of sequential equilibrium. These refinements involve more sophisticated reasoning about interpretation of actions, especially those off the equilibrium.

The first refinement is based on the iterated elimination of dominated strategies; this is closely connected with the idea of rationalizability.[21] Alternative restrictions on beliefs are provided by universal divinity[22] and perfect sequential equilibrium.[23] These three alternatives to sequential equilibrium are discussed in turn.

One can judge the merits of these approaches in either of two ways: by their axioms or by their results. A comparison of the axioms shows three plausible restrictions on what might constitute a reasonable belief. As we move from elimination of dominated strategies to universal divinity to perfect sequential equilibrium, the restrictions employ increasing levels of sophisticated logic. A comparison of the results reflects this ranking. Each increased level of sophisticated reasoning further reduces the scope for self-fulfilling expectations to support a no-intervention equilibrium. Therefore, if one intuitively feels that this equilibrium is

[20] Note that if $a = 1/2$ and $x > 1/4$, then the model predicts that a country with $c = 0$ would strictly prefer to intervene. Hence the reversed equilibrium is ruled out.

[21] David Pearce, "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica* 52 (July 1984), 1029–50; B. Douglas Bernheim, "Rationalizable Strategic Behavior," *Econometrica* 52 (July 1984), 1007–28.

[22] Jeffrey Banks and Joel Sobel, "Equilibrium Selection in Signaling Games," *Econometrica* 55 (May 1987), 647–62.

[23] Sanford Grossman and Motty N. Perry, "Perfect Sequential Equilibria," *Journal of Economic Theory* 39 (June 1986), 97–119.

unreasonable, then the refinements provide a more general framework for characterizing just what makes it unreasonable. But one may well take the other view. There are circumstances where the no-intervention equilibrium seems reasonable (such as when there are no objective benefits from intervention) and yet it is ruled out. From this second perspective, the results present a specific example that questions the merit of the axioms. This paper does not attempt to take sides. The different approaches are presented in a way that leaves it to the readers to form their own opinions.

Elimination of Dominated Strategies

A strategy is dominated if there is some other strategy that leads to a higher payoff no matter how the action is interpreted.[24] The elimination of dominated strategies makes the presumption that a player will not follow a dominated strategy. In the present context this says that it is unreasonable to believe that a country should ever intervene if that action lowers its payoff (relative to not intervening) given *any* possible expectation about the unobserved cost parameter following intervention. And conversely, it is unreasonable to believe that a country should ever fail to intervene if nonintervention yields a lower payoff given *any* possible expectation about the unobserved cost parameter following nonintervention. In both cases, we eliminate dominated strategies from consideration.

To form beliefs about who will do what, we restrict attention to countries with a range of unobserved costs for which the proposed action is not dominated (if such a cost exists). The procedure is then iterated as illustrated below. With each cycle our beliefs are increasingly restricted as to who may take any action. Sometimes the beliefs will converge on a unique solution, thus selecting one of the sequential equilibria. More generally, we are able to show that some of the earlier reversed sequential equilibria are no longer sustainable with this restricted set of beliefs.

What are the possible beliefs? For the action that everyone takes, the posterior belief must equal the prior belief ($c = 1/2$). The worst belief about a country that fails to follow the proposed equilibrium behavior is that its $c = 1$. Hence, the changed perception of c following intervention is at most $1/2$ and the best possible net payoff to intervening is then

$$x - c + a/2.$$

If this is negative, intervention is dominated by not intervening. In our numerical example, $x = 1/4$ and $a = 1/2$. Thus intervention is domi-

[24] Note that this domination is over other strategies, not over an opponent.

nated for all countries with $c > 1/2$. No matter what others think about the reputation of noninterveners, it is not worthwhile to intervene whenever $c > 1/2$.

This simple idea allows us to eliminate the (reversed) sequential equilibrium that results in no-intervention. The reason is that we may no longer suppose that an intervener has cost 1. A country with cost 1 finds intervention to be a dominated strategy. The worst reputation we can assign to an intervener is to believe that its cost is 1/2—this is the highest cost for which intervention is not strictly dominated.

To see whether an equilibrium in which no one intervenes is still possible, we must also consider the other side of the cost-benefit ledger. What is the most favorable reputation that can be given to a nonintervener? Because intervention is a dominated strategy, all countries with costs from (1/2, 1] must be counted as noninterveners. Therefore, the best reputation we can give to a nonintervener is to add all the types [0, 1/2) to the list of noninterveners. This gives an expectation of 1/2 for those who do not intervene. These beliefs are illustrated in Figure 1.

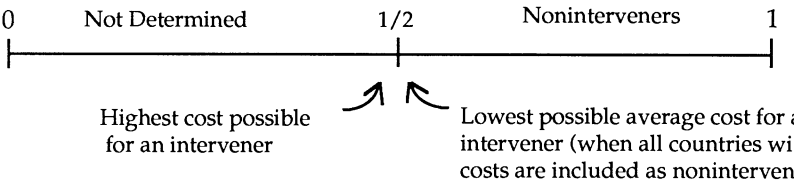

FIGURE 1

To illustrate the iterated elimination of dominated strategies, we now redo the cost-benefit calculations. The worst we can believe about an intervener is that its cost is 1/2; the best we can believe about a nonintervener is also 1/2. Thus when reputations are chosen to be maximally biased against intervention, we find that interveners and noninterveners are given identical reputations. The chosen action does not affect reputation. Hence all countries with $c < x$, or costs in the range [0, 1/4) in our numerical example, will find intervention a dominant strategy.

This iterated elimination of dominated strategies limits our ability to use self-fulfilling expectations to punish those who intervene and reward those who do not. We must now expect that all those with costs below 1/4 will intervene while those with costs above 1/2 will not. As illustrated in Figure 2, our only freedom is to choose beliefs about those in the interval [1/4, 1/2].

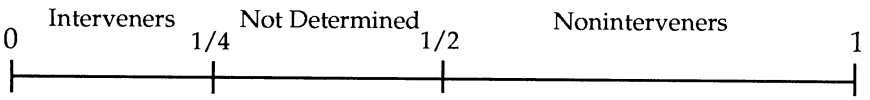

FIGURE 2

The worst reputation for an intervener arises when we presume that all countries with unobserved costs from [0, 1/2] will intervene. The best reputation from the nonintervener's perspective follows from the belief that all types from [1/4, 1] will not intervene. (Note that these beliefs are not internally consistent.) Reapplying the cost-benefit calculations shows that the reputation effect is now worth 3/8 so that countries with $c < 7/16$ intervene whereas those with $c > 1/2$ still find intervention a dominated strategy. The range of ambiguity is reduced even further, down to the countries with costs in the interval [7/16, 1/2].

How we continue from here gets more complicated. But the limit should also be apparent. There is only one equilibrium that is left after we iteratively eliminate dominated strategies. The equilibrium is the one in which countries with costs less than 1/2 intervene and those with costs above 1/2 do not. This is the predicted outcome from the original separating equilibrium as $c^* = x + a/2 = 1/2$.

The fact that elimination of dominated strategies leads to a unique equilibrium makes that outcome particularly appealing. But even if the process did not converge to a unique solution, the usefulness had already been demonstrated: for the case with $x = 1/4$ and $a = 1/2$, the reversed equilibrium in which no one intervenes is not consistent with the iterated elimination of dominated strategies. Although a weak country may claim that its restraint is a sign of strength, this is not a reasonable interpretation given the parameters above. Here, a reasonable interpretation means that we exclude the possibility that a country follows a dominated strategy. This restriction on what is reasonable to believe is an effective tool for limiting the scope of self-fulfilling expectations.

UNIVERSAL DIVINITY

A stronger restriction on beliefs is called universal divinity.[25] Universal divinity focuses on the consistency condition 2(b). Under a sequential equilibrium one is free to assign an arbitrary reputation to a country that takes an action that should never occur. But is it reasonable to believe that a country with high unobserved intervention costs would make an

[25] Banks and Sobel (fn. 22).

unexpected intervention while one with low unobserved intervention costs would not? Universal divinity classifies this as an unreasonable belief. We should not assume that a country that loses a lot by taking the unexpected action has deviated while those who would lose less do not. Instead, the argument is made that the expectation of who might have taken the out-of-equilibrium action should include all countries with negative payoffs whose losses are less than some cutoff amount. This idea has power to break a (reversed) sequential equilibrium in which no one is supposed to intervene; it has no power to break a sequential equilibrium in which everyone is supposed to intervene. We look at these two self-fulfilling equilibria in turn.

In an equilibrium where no country should intervene, the cost-benefit calculations must indicate a negative payoff to any country that intervenes.[26] Yet upon seeing an intervention, which country with a negative payoff is supposed to have taken this unprofitable action? Universal divinity argues that it is unreasonable to assume that those who stand to gain the least from intervention will do so without also believing that everyone else will deviate. Specifically, it is unreasonable to suppose that the intervener is a country with payoff -0.8 but not a country with payoff -0.4. One must choose some lower endpoint and take the average over countries with net payoff no worse than that amount. In our model this is equivalent to believing that the intervener has unobserved costs less than some critical c. The worst reputation for an intervener is to suppose that the critical c is as high as possible; everyone intervenes from cost $c = 1$ on downward. In this case, $c_i = 1/2$. This is also the reputation given to a nonintervener, $c_n = 1/2$. Since no one is supposed to intervene, a nonintervener's reputation must remain unchanged from its prior level of 1/2. Thus intervention does not improve reputation in this best-case scenario for nonintervention.

For the example with $x = 1/4$ and $a = 1/2$ and beliefs that accord with universal divinity, we cannot support an equilibrium where no country intervenes. A country with zero (or even close to zero) intervention costs is willing to intervene provided this does not hurt its reputation too much. Since we are not allowed to suppose that interveners have higher costs than noninterveners, there is at worst no loss in reputation. Hence, the cost-benefit calculation become strictly positive in favor of intervention for low-cost countries ($c < 1/4$). Just as in the case of iterated elimination of dominated strategies, we have broken the sequential equilibrium in which nonintervention is viewed as a sign of strength.

[26] Otherwise, intervention must be expected.

Universal divinity has no power to disrupt the sequential equilibrium in which everyone is presumed to intervene. In this case, the cost-benefit calculations must come out negative against those who might fail to intervene. The country for which this comparison is the least negative is the one with the highest intervention cost ($c = 1$). Hence, universal divinity allows us to count as reasonable the belief that a nonintervener has $c = 1$. There is no one else who finds nonintervention relatively more favorable and who thus must be added to the list. This implies a loss of reputation from 1/2 down to 1 for a country that fails to intervene. If reputation were important enough (a is large) then all countries would choose to intervene. For the cases with $c^* = x + a/2 \geq 1$, there remains a self-fulfilling equilibrium in which everyone intervenes.

Perfect Sequential Equilibrium

A third approach is called perfect sequential equilibrium[27] or credible neologisms.[28] The idea is to look for an internally consistent set of beliefs about who might have deviated. Is there a range of costs such that if we believe that countries with those costs deviate then it is in their interest to do so and it is not in the interest of any of the other countries to deviate? If such an internally consistent belief exists, then the original equilibrium fails this third refinement test.

An example illustrates the technique. Consider a situation in which x is negative. Since it is common knowledge that there is no objective benefit from intervention, the preferable outcome is that no one intervenes. For the case with $x = -1$ and $a = 3$, there is such an equilibrium. The no-intervention equilibrium is supported by $c_n = 1/2$ and $c_i = 1/2$. Since there is no gain in reputation and intervention has an observed negative payoff, there is no reason for any country to intervene.

This solution satisfies elimination of dominated strategies. The best possible payoff for intervention is $x - c + a/2 = 1/2 - c$. Thus all countries with $c > 1/2$ find intervention a dominated strategy. This implies that the lowest value of c_n is 1/2 and the highest value of c_i is also 1/2 (see Figure 1). But these are exactly the reputations used to support the prediction that no one should intervene. Unlike the earlier example, the group of countries for which $x > c$ is the null set and so there is no country that must now be classified as an intervener. There is no scope for iteration and the no-intervention equilibrium remains.

[27] Grossman and Perry (fn. 23).

[28] Joseph Farrell, "Credible Neologisms in Games of Communication" (Mimeo, University of California, Berkeley, 1984).

This solution also satisfies universal divinity. The requirement concerning beliefs about the zero-probability event is that we include all countries whose payoffs for that action are no worse than some amount. When intervention is the zero-probability event, then the lowest reputation arises when everyone is included (since the countries with low unobserved intervention costs must be included whenever high-cost countries are included). This leads to $c_n = c_i = 1/2$, which are the numbers used to support the no-intervention solution.

The point of this exercise is to show that the no-intervention outcome does not satisfy perfect sequential equilibrium (PSE). Whether this leads one to think the no-intervention outcome is less reasonable or PSE is less reasonable is left for the reader to decide. Under PSE, a country that observes the zero-probability event of an intervention must try to justify its occurrence. It is possible to do so using the following beliefs. One imagines that countries with unobserved costs from 0 to 1/2 will intervene ($c_i = 1/4$) and those with costs from 1/2 to 1 will not ($c_n = 3/4$). With these beliefs, the reputation gain from intervention equals 1/2, which is valued at 3/2. Comparing the unobserved cost of intervention with the observed payoff, we find that the beliefs are confirmed:

$$c^* = x + a/2 = -1 + 3/2 = 1/2.$$

Countries with $c < 1/2$ prefer to intervene and those with $c > 1/2$ prefer not to.

To put this example back into the definition of PSE, we have found a range of costs [0, 1/2) such that if we believe that countries with those costs deviate (by intervening) then it is in their interest to do so and it is not in the interest of any of the other countries to deviate. Since such an internally consistent belief exists, the original no-intervention equilibrium fails the PSE test.[29]

Upon reflection, one might choose to argue that this separating equilibrium is the more sensible outcome. It is the outcome predicted by the base-case model for sequential equilibrium. Even though the average country ends up worse off, the countries with low unobserved costs benefit at the expense of those with high costs. The reputation of interveners is better by 1/2, and with the large weight placed on reputation this justifies paying the observable intervention costs, x.[30]

[29] Perfect sequential equilibrium is not a panacea. It is much harder to show that something satisfies PSE than to show that it does not. In fact, one of its failings is that it is possible that no equilibrium will satisfy this test. In our example this problem does not arise as the separating equilibrium satisfies PSE.

[30] As this intuition suggests, it is not the case that PSE eliminates all no-intervention equilibrium. If x is negative and a is small, the no-intervention equilibrium satisfies PSE.

This separating outcome may seem even more appealing if we examine the intuition for why the no-intervention equilibrium fails PSE. The problem with expecting no one to intervene is that if we do observe the unexpected, this "mistake" can be justifed if we presume that the intervener was a low-cost country ($c < 1/2$). But if we are willing to give this reputational advantage to a country that breaks the equilibrium by intervening, for the sake of consistency we should penalize those who fail to intervene ($c > 1/2$). The advantage of these expectations is that they offer a consistent explanation for the observed behavior.

VI. Conclusions

The potential for a reputation paradox arises in a world with imperfect information. Because unobserved motivations matter in predicting behavior, there is a need to make presumptions about the unobserved costs and benefits of any action. This is particularly difficult when it is an action that cost-benefit calculations predict should never occur. The endogenous relationship between the presumptions and the predictions complicates the inference problem. The predictions are based on the presumptions, but the presumptions must be consistent with the predictions. When the predictions suggest that something should not occur, how are the presumptions to be made consistent?

Game theory offers (at least) four possible answers. We use the setting of a rational deterrence model to explain and motivate these recent refinements in equilibrium theory. The refinements illustrate what is paradoxical about an equilibrium in which no country intervenes because intervention is associated with weakness rather than strength. The most sophisticated refinement can explain away the no-intervention outcome even when it seems the more natural case, such as when the intervention payoffs are all negative. The advantage of our stylized model is that it is possible to follow the connection from assumption to conclusion. Here one may choose among refinements by seeing the results. As with Christmas pudding, the proof is in the eating.