National Security Strategy: Sequential Equilibrium

Professor Branislav L. Slantchev

January 1, 2014

Overview We have now defined the concept of credibility quite precisely in terms of the incentives to follow through with a threat or promise, and arrived at a solution concept of perfect equilibrium which takes it into account. We now turn to the question of incomplete information and study an escalation game in which the defender is uncertain about whether its opponent is resolute or not. We find that we have to consider not just strategies but beliefs, and refine our solution concept to account for both, calling it sequential equilibrium. The solutions to this game uncover some rather surprising dynamics of deterrence and compellence.

We have now defined the concept of credibility very precisely. A threat (or a promise) is not credible if the player would not carry it out if given a choice. We then used this idea to argue that a reasonable solution to a game should not depend on a player using incredible threats because his opponent would never believe them, and would therefore ignore them when determining her own best response. We introduced the perfect equilibrium solution concept that rules out Nash equilibria which depend on such unreasonable behavior. Further, we studied an easy way to analyze complete information games with backward induction.

The solutions to the two escalation games, one with a weak challenger and the other with a tough challenger, demonstrated that the perfect equilibria are very different. In the first case, the challenger did not have a credible threat to attack, so the defender had a credible threat to resist, which in turn was sufficient to deter the challenger from escalating in the first place. The perfect equilibrium outcome was the status quo. In the second case, where the challenger could credibly threaten to launch an attack if resisted, the defender was compelled to concede, which in turn implied that she would fail to deter the opponent from escalating in the first place. The outcome was capitulation by the defender.

In both cases, the probability of war in equilibrium was zero, which makes intuitive sense. If everything in the game is common knowledge, then players would succeed in avoiding the costly confrontation. Note in particular that even though the resolute challenger prefers the status quo to war, he still escalates because he knows that the defender will back down. Furthermore, if the defender knows that the challenger will capitulate, she will resist and her threat will work even though she is weak.

Although very useful to illustrate the idea of credibility, these models actually pose more questions than they answer. In the real world, it is very likely that adversaries will not know the resoluteness of the opponent. So, what would happen if this is the case? Further, from our simple simultaneous move crisis game we know that a little uncertainty can immediately generate a positive probability of war in the mixed strategy equilibrium. Yet war is sure not to occur in the perfect equilibria of the escalation models. We now turn to the analysis of an escalation game under incomplete information.

1 The Escalation Game with Incomplete Information

We have seen how to model games of incomplete information as games of imperfect information. A brief review is in order. Suppose that the weak defender, D_W, does not know whether the challenger is tough, C_T, or weak, C_W. (Her weakness is common knowledge.) The defender does have some prior belief (e.g. from previous interactions, from results of CIA analysis, etc.) that the probability that the challenger is tough is $p \in (0, 1)$. Of course, we could assign a specific probability

to p, but we prefer to conduct the analysis for arbitrary values of the prior beliefs so we can apply the results to all sorts of situations. That is, we want to be able to say things like "if D's prior belief is pessimistic (that is, it assigns a high probability to the challenger being tough), then the equilibrium would be such and such, and if D's prior belief is optimistic, then the equilibrium would be so and so." The idea is to make our results as general, and therefore useful, as possible.

Recall that to model the uncertainty about the challenger's type, we introduce the fictitious player Nature, N, which "chooses" the tough type with probability p, and the weak type with probability $1 - p$. The challenger knows his type when making his move, but the defender can only observe the move and does not know which type of opponent actually made it. The situation is represented in Figure 1.
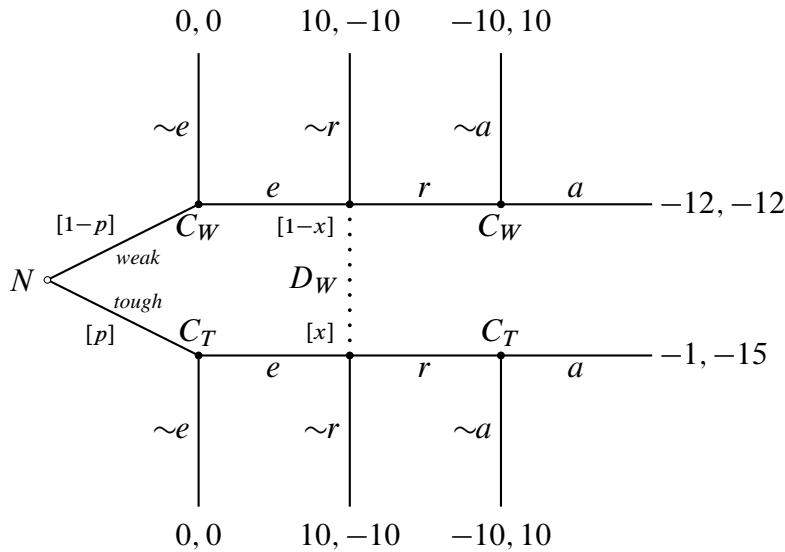

Figure 1: Escalation Game with a Weak Defender and Incomplete Information about the Challenger.

The information set for player D contains two nodes, one following escalation by C_T and another following escalation by C_W, because even though D can observe escalation, she does not know wether the tough or the weak challenger was responsible for it. The label x represents D's **belief** that it was the tough one who escalated, and $1 - x$ represents her belief that it was the weak one who escalated. In other words, x is D's estimate that the challenger is tough given that escalation has occurred. We shall see shortly how D will calculate this belief. For now, all we need to keep in mind is that because D is unsure about the nature of her opponent, she may be unable to predict what he will do when resisted. From her perspective, resistance will lead to war if C is tough but to peace (with victory) if C is weak. Since D does not like war, she has to figure out if the risk of resistance is worth it. Estimating this risk depends on what she believes about the type of her opponent. Intuitively, this belief, x, should depend on her priors and on any new information

she can glean during the crisis itself (i.e., from the challenger's decision to escalate).

We can begin solving this game by backward induction. At the last node for the weak type, C will never attack because attacking yields -12, while not attacking yields -10. Therefore, in any perfect equilibrium, the weak type would capitulate if resisted. On the other hand, at the last node for the tough type, C will always attack because doing so yields -1, while not attacking yields -10. Therefore, in any perfect equilibrium, the tough type would attack if resisted. We fold back the game by removing the branches representing actions that are not credible (attack for the weak and capitulation for the tough) because these can never occur in a perfect equilibrium.

Because resisting the weak type results in capitulation by the challenger and resisting the tough type results in war, we replace C's last decision nodes with the payoffs for the outcomes that would result in these nodes are ever reached by D's resistance. The result is shown in Figure 2. Note that, as our intuition suggested, D's choice to resist can lead to two different outcomes depending on the type of opponent she faces.
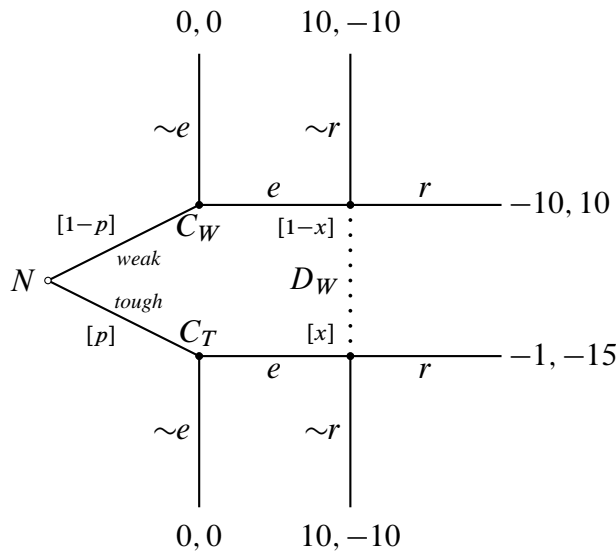

Figure 2: The Escalation Game After Pruning the Last Nodes.

Unfortunately, we cannot continue the backward induction. Why? Because the optimality of D's action depends on what she thinks about C; that is, whether D believes that she is at the lower or upper node in her information set. For example, if D knew that her opponent were weak (she would be at the upper node), her belief would be $1-x = 1$, or $x = 0$. In this case, she would prefer to resist because doing so would yield 10, while playing $\sim r$ would yield only -10. However, if D believed that C were tough (she would be at the lower node), her belief would be $x = 1$. In this case, she would prefer not to resist because doing so would yield -10, while playing r would yield -15. As we expected, the optimal action crucially depends

on this belief. We now formalize this idea by analyzing how D's optimal behavior depends her beliefs.

2   Sequential Rationality

We shall call a player's strategy **sequentially rational** if it is a best response to the opponent's strategy given the player's beliefs. (For now, we do not specify where these beliefs come from.) A strategy is sequentially rational for some belief x if the player would actually want to play this strategy if his belief ever became x. This generalizes the notion of best response by explicitly taking into account the beliefs that the player has about his opponent's behavior.

Returning to our example, let's determine the beliefs that would make r a best response by D, and the beliefs that would make $\sim r$ a best response. That is, we are asking the question, "What could D believe about the type of her opponent that would make resistance a best response?" To put it another way, we want to find the belief that rationalizes a particular strategy.

Note that at her information set, D only has two choices: resist or not. Let's calculate the expected utility from resisting, like we did for the mixed strategies before. If player D chooses r, then she will either end up at war if C is tough or victory if C is weak. She believes that C is tough with probability x, so from her perspective resistance leads to war (payoff of -15) with probability x and victory (payoff of 10) with probability $1 - x$. The expected payoff from resistance is then:

$$\mathrm{U}_D(r) = x(-15) + (1 - x)(10) = 10 - 25x.$$

If player D chooses $\sim r$, then the outcome will be her capitulation regardless of the type of opponent. We could write out the expected payoff: she would get -10 with probability x and -10 with probability $1 - x$, or:

$$\mathrm{U}_D(\sim r) = x(-10) + (1 - x)(-10) = -10.$$

When would she choose r? When the expected utility from doing so exceeds the expected utility of choosing $\sim r$:

$$\mathrm{U}_D(r) > \mathrm{U}_D(\sim r)$$
$$10 - 25x > -10$$
$$x < {}^{20}\!/_{25} = {}^{4}\!/_{5} = 0.8.$$

This gives is the *critical threshold* for the belief that rationalizes resistance. If D came to believe that C is tough with probability less than 80%, then the rational thing to do will be to resist. There is risk in this action: after all, the challenger *could* turn out to be tough, and in that case resistance will cause war. However, the risk is worth it given her beliefs. Optimism (belief that one's opponent is weak)

can generate a risk of war, as many historians and political scientists have noted. Pessimism, on the other hand, may lead to peace. If D's estimate of the chances of C being tough goes above 80%, then the risk of war becomes intolerable, and she will submit. In this case, her belief causes her to think that resistance is too likely to lead to an attack because the challenger is very likely to be tough. Given her aversion to war, D will not run this risk.

In our terminology, the strategy r is sequentially rational if $x < 0.8$. That is, the strategy of resisting is sequentially rational if, and only if, D believes that C is tough with probability less than 80%. This means that the strategy $\sim r$ is sequentially rational if $x > 0.8$. That is, the strategy of not resisting is sequentially rational if, and only if, D believes that C is tough with probability greater than 80%. Finally, if $x = 0.8$, then D is indifferent between the two strategies. As before, this means that they are both best responses; both r and $\sim r$ are sequentially rational. As we know, if this is the case, then D can mix between them, so she can play r with some probability q and $\sim r$ with some probability $1 - q$, where $0 \leq q \leq 1$.

We shall express D's pure strategies in terms of this mixed strategy. That is, $q = 1$ is the same as the pure strategy r, and the mixed strategy $q = 0$ is the same as the pure strategy $\sim r$. To summarize our findings, D's sequentially rational best responses are:

$$\text{BR}_D(x) = \begin{cases} q = 1 & \text{if } x < 0.8 \\ q = 0 & \text{if } x > 0.8 \\ 0 \leq q \leq 1 & \text{if } x = 0.8. \end{cases}$$

Notice that these best responses are now functions of D's *beliefs*. Sequential rationality critically depends on beliefs: an action is only sequentially rational *given some beliefs*. One cannot evaluate its optimality without considering them. But where do these beliefs come from?

3 Consistent Beliefs

When we think about D's belief x (the probability she assigns to the opponent being tough), we intuitively know that it should depend on two things: (i) the initial belief D had before C escalated, and (ii) the fact that C actually did escalate. That is, x is going to be somehow related to the information D had before the crisis, and the new information acquired during the crisis from observing her opponent's behavior and making inferences about what could have produced such behavior.

We have already decided that prior to the crisis, D's belief is represented by the move by Nature. That is, this chance move was designed to convey the idea that D believed that C was tough with probability p, and weak with probability $1 - p$. We shall call this p, player D's **prior belief** for obvious reasons. As we discussed, this belief could come from prior experience with the challenger, or analysis of challenger's behavior in other crises, or analysis by experts (this is what the CIA,

army intelligence, and a host of other organizations actually do), or even impressions from C's interactions with other players (we shall see quite a bit of that when we go over historical cases). At any rate, this prior belief p exists before the game begins.

If the challenger escalates, then D must take this into account and revise her belief accordingly. Why? Because the challenger knows his own type, his choice of strategy will depend on that *private information*. But if that's the case, then the defender can look at the choices C is making and perhaps *infer* what type of opponent she faces. The defender will attempt to learn this private information so that she may choose her best response accordingly. Obviously, the challenger knows that she will do this and will try to *manipulate* this belief in order to *induce* a response that he likes best. Of course, the defender knows that the challenger knows what she is doing, so she will take into account his attempt to manipulate her beliefs when she makes her inferences, and so on. The question then is: how does D **revise her prior belief** in the light of the new information conveyed by escalation?

What we are asking is how to compute $x = \Pr(C_T|e)$, which reads "what is the probability that the challenger is tough given that he escalated?" We call x the **posterior belief** (or updated belief) because it takes into account the information that C has escalated. Of course, D cannot just arbitrarily interpret escalation: the information provided by this move must be **consistent** with what constitutes rational behavior by C. For example, if the weak type would never escalate in equilibrium, then upon observing escalation D should never believe that she might be facing the weak challenger.

This means that D has to take into account the challenger's strategy when making her inferences. Since we allow for mixed strategies, when D updates her belief she will note the *probability* that the tough challenger would escalate and the *probability* that the weak one would escalate. Since we do not know these probabilities yet, let α denote the probability that C_T chooses e, and let β denote the probability that C_W chooses e. Recall that player C has two information sets in our revised game in Figure 2. Therefore, his strategy should include two components: what to do if he is the tough type, and what to do if he is the weak type. A pure strategy would be $(e, \sim e)$, which says "escalate if tough, do not escalate if weak." There are four type-contingent pure strategies.[1]

With α and β, we are just writing the mixed strategies, so (α, β) is the type-contingent mixed strategy which says "escalate with probability α if tough, and escalate with probability β if weak." Of course, this means also "do not escalate

---

[1]Strictly speaking, the strategy must include the action to take after D's choice to resist. We already know that subgame-perfection requires the tough challenger to attack and the weak to capitulate. Hence, any equilibrium we find must specify these actions as part of the optimal strategy for the challenger. To reduce clutter, I will not write them explicitly but instead focus on the initial choice to escalate.

with probability $1 - \alpha$ if tough, and do not escalate with probability $1 - \beta$ if weak." For example, the mixed strategy $(1, 0)$, which denotes $\alpha = 1$ (tough type escalates with certainty) and $\beta = 0$ (weak type does not escalate with certainty) is the same as the pure strategy $(e, \sim e)$. The mixed strategy $(0.5, 0.3)$ would be read as "escalate with probability 0.5 if tough, and escalate with probability 0.3 if weak." To keep things clear, the revised Figure 3 labels the branches with their corresponding probabilities.
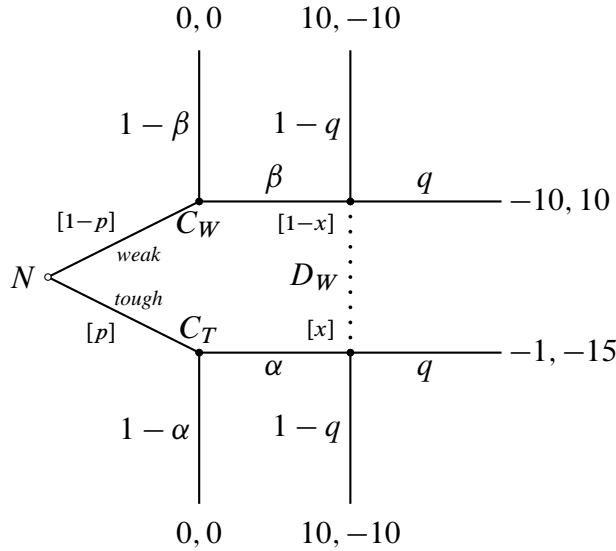

Figure 3: The Escalation Game With Mixed Strategies.

Note that D's mixing probability q must be the same for both nodes in her information set because she cannot condition her behavior on C's type if she does not know it. On the other hand, C's mixing probabilities at his two nodes can be different because they are in different information sets: he does know his type and can therefore condition his behavior on it.

How do we calculate the posterior probability x given the prior probability p and C's mixing probabilities α and β? There is a simple formula that allows us to compute the posterior belief from the prior belief and the new information. It is called **Bayes rule**, which some of you may have seen in elementary courses on probability theory or statistics. In our case, this rule allows us to answer the question: Given that C has escalated, what is the probability that C is tough? Intuitively, to answer this question, we need to figure out what the probability of escalation is in the first place. Knowing that, we can estimate what portion of that probability "belongs" to the event that escalation was caused by a tough challenger.

Escalation can be caused by either type of challenger. Because the two types are mutually exclusive (if the challenger is weak he cannot be tough) and exhaustive (there are only these two possible types of challenger), the probability of escalation is simply the sum of the probability that the tough type escalates and the probability

that the weak one does. The tough type escalates with probability $\Pr(e|C_T) = \alpha$ and the weak one with probability $\Pr(e|C_W) = \beta$. We read the expression $\Pr(e|C_T)$ as "the probability that the challenger escalates if he is tough." The probability that the challenger is tough *and* escalates is then $\Pr(C_T, e) = \Pr(C_T) \times \Pr(e|C_T) = p\alpha$. This is different from the *conditional* probability $\Pr(e|C_T)$, which assumes that the challenger is, in fact, tough. The *joint* probability of type and escalation takes into account the uncertainty about the type. In other words, whereas the conditional probability measures how likely escalation is if the challenger is tough, the joint probability measures how likely it is for the challenger to be tough and to escalate. The probability that the challenger is weak *and* escalates is $\Pr(C_W, e) = \Pr(C_W) \times \Pr(e|C_W) = (1-p)\beta$.

Since D is uncertain about the type she faces, from her perspective the probability of escalation is $\Pr(e) = \Pr(C_T, e) + \Pr(C_w, e) = p\alpha + (1-p)\beta$. This quantity is the *total probability* of escalation. Now that we know how likely escalation is in the first place, we can compute the chances that it was caused by the tough challenger. We want to know $x = \Pr(C_T|e)$, that is, the *conditional* probability that C is tough given that escalation has occurred. But this is simply the probability that C is tough and escalates divided by the probability that escalation occurs, $\Pr(C_T, e) / \Pr(e)$, or:

$$x = \Pr(C_T|e) = \frac{\Pr(e|C_T)\Pr(C_T)}{\Pr(e|C_T)\Pr(C_T) + \Pr(e|C_W)\Pr(C_W)} = \frac{p\alpha}{p\alpha + (1-p)\beta}.$$

This is Bayes rule. We require that D update her beliefs using this formula. The only posterior beliefs that we shall consider reasonable are ones that are derived from the prior beliefs and the strategies by applying Bayes' rule, if possible. When beliefs are computed with this formula (which takes into account the strategy of the opponent), we say that **beliefs are consistent with the strategies**.

By "if possible" I mean "whenever the formula is defined." Note that if $\alpha = \beta = 0$, then the formula is not defined because one cannot divide by zero. In other words, one cannot condition on zero-probability events in this way. Thus, if no type of C escalates ($\alpha = \beta = 0$), then escalation is a zero-probability event and should not occur. What is D to believe if this event actually does occur? This is an open question and a people are still trying to figure out what a "reasonable" belief should be in this case. For example, we all expect the sun to rise in the east, so the sun rising in the west is a zero-probability event. What would you believe if one day you woke up and the sun was rising in the west? For our purposes, it is sufficient to assume that if the formula is not defined, then any belief is consistent with the strategies. That is, we can assign whatever beliefs we wish.

Let's see how the formula works. Suppose C's strategy is $(e, \sim e)$; that is, the tough one escalates with certainty, and the weak one does not, also with certainty. In our mixed-strategy notation where the strategy is denoted by (α, β), it translates into $(1, 0)$. What should x be? Intuitively, we think that $x = 1$ should be the result because if escalation does occur and only the tough one escalates, the posterior

belief after escalation should be that D is facing the tough one for sure. This is indeed the case: $x = \frac{p(1)}{p(1)+(1-p)(0)} = 1$, as expected. Note that it does not matter what p is in this case.

Suppose now that C's strategy is $(0, 1)$; that is, the tough one never escalates, but the weak one always does. Then $x = \frac{p(0)}{p(0)+(1-p)(1)} = 0$. That is, after observing escalation, D would conclude that C is weak for sure. This is also intuitive and also does not depend on p. In both instances, the prior belief is irrelevant because C's strategy must lead to certain inferences.

Observe that it is quite possible to obtain certain inferences even if C plays a partially mixed strategy. For example, suppose $\alpha \in (0, 1)$ and $\beta = 0$; the tough type escalates with some positive probability and the weak type never does. Again, $x = \frac{p\alpha}{p\alpha+(1-p)(0)} = 1$, so the inference depends neither on the prior nor on the precise mixing probability by the tough type.

Of course, things are not so simple if $\alpha = 1$ and $\beta \in (0, 1)$. Here, the tough type escalates for sure and the weak escalates with positive probability. Escalation is no longer a sure signal of C's type. Suppose, for the sake of illustration, that $p = \frac{1}{2}$ and $\beta = \frac{1}{3}$. Then Bayes' rule yields:

$$x = \frac{p(1)}{p(1) + (1 - p)\beta} = \frac{(\frac{1}{2})(1)}{(\frac{1}{2})(1) + (\frac{1}{2})(\frac{1}{3})} = \frac{3}{4} = 75\%.$$

In other words, if D had this prior and thought that C played this particular strategy, her consistent belief following escalation would be that the challenger is tough with 75% probability. Whereas D will still be uncertain about the type of her opponent, she would have learned something from his escalation. Recall that she began the game believing that the chance of C being tough was 50%. Following escalation, she revises her belief upward and now estimates that this chance is 75%. This makes intuitive sense: the challenger's strategy is such that the tough type escalates with a higher probability than the weak type. We would expect this to cause D to revise her estimate upward when escalation does occur. Bayes' rule gives us the precise result of this intuitive revision.

4 Sequential Equilibrium

We now put together the ideas of **sequential rationality** and **consistent beliefs** to refine our solution concept to take them into account. A strategy profile is a **sequential equilibrium** if the strategies for all players are sequentially rational and beliefs are consistent with these strategies. This is a generalization of the perfection requirement in that it takes into account beliefs explicitly. Our search for mutual best responses is now a bit more complicated because we have to consider not just the strategies but also the accompanying beliefs in our solutions. After all, we know that beliefs rationalize strategies but that the strategies themselves are used to derive these beliefs. We have to solve for the combination simultaneously.

10

Let's proceed with our example. We now know how D is going to update her beliefs for any strategy that C might play. We further know how D is going to behave given these beliefs. The only remaining question is how C would play when he knows that his action is going to influence D's beliefs. Recall that D will use Bayes rule to make consistent inferences from C's strategy. C knows that and will attempt to pick strategies that induce inferences he prefers. For example, if he could get D to believe that he is tough with sufficiently high probability (any $x > 0.8$), then D would rationally respond by capitulating. It is in C's interest to attempt to manipulate D's beliefs to cause her capitulation. Conversely, C really does not want D to believe that he is weak with high probability (any $x < 0.8$) because if she ever did acquire this belief, she would resist. Hence, the challenger (and in particular the tough type) really wants to prevent D from making this inference. Of course, D is perfectly aware of these incentives and knows that C will try to manipulate her beliefs. We now see how all of this resolves itself in equilibrium.

4.1 Separating Equilibria

We first consider the four pure strategies for C. Suppose C plays the strategy $(\alpha = 1, \beta = 0)$; that is, escalate if tough, do not escalate if weak. In this case, D's updated belief will be, by applying the formula, $x = 1$. From D's best response function, we know that $\mathrm{BR}_D(1)$ is $q = 0$, so her best response to this belief is to capitulate. But is C's strategy a best response to capitulation? Let's compare the expected utility of the weak type who is supposed to stay with the status quo with certainty. If he does not escalate, his payoff is $\mathrm{U}_{C_W}(\beta = 0) = 0$ because the status quo prevails. If he deviates and escalates instead, D (wrongly thinking escalation was caused by the tough type) will back down, and the expected payoff is $\mathrm{U}_{C_W}(\beta = 1) = 10$. That is, the weak challenger can definitely do better by escalating. But in equilibrium no player should have an incentive to deviate from his strategy. Thus, the strategy $(1, 0)$ cannot be a part of any sequential equilibrium. This makes intuitive sense. If D were to believe with certainty that C is tough, she would always back down. But precisely because she would back down, the weak challenger will do better by changing strategy and escalating. After all, there will be no risk in having to capitulate himself.

Suppose now that C plays the strategy $(0, 1)$; that is, do not escalate if tough, escalate if weak. In this case, D's posterior belief will be $x = 0$, so her best response will be to resist ($q = 1$). Is C's strategy then a best response to this? It is not: The weak type's expected payoff from escalating is, given that D will resist, $\mathrm{U}_{C_W}(\beta = 1) = -10$, which is worse than the expected payoff from not escalating which is $\mathrm{U}_{C_W}(\beta = 0) = 0$. Therefore, this strategy cannot be a part of any sequential equilibrium. This also makes intuitive sense. The only way D would believe with certainty that C is weak following escalation is if only C_W escalated. But given this belief, D's best response is to resist, which C_W wants to avoid in the

first place, so C_W will never escalate, which in turn implies that there is no way for D to hold this belief.

The strategies $(1, 0)$ and $(0, 1)$ are called **separating** because the different types of C choose different actions with certainty. That is, the types "separate" themselves by their actions. Of course, when C plays a separating strategy, D can infer with certainty what C's type is, as we have already seen. A sequential equilibrium, in which players play separating strategies is called a **separating equilibrium**. As we have seen, there are no separating equilibria in this game. It is not reasonable to expect that C will choose a strategy that would reveal to D whether he is weak or tough.

This is a crucially important result. Think about what it means. If C played a separating strategy in equilibrium, D would either capitulate for sure (because she believes C is tough) or resist for sure (because she believes that C is weak). But as we know from the complete information case, if she resists the weak type, the weak C never escalates in the first place. Thus, if C plays a separating strategy in equilibrium, either C never escalates (status quo) or D capitulates. In other words, the probability of war would be zero, just like in the complete information case.

If there were any way for C to reveal his type to D, war would be avoided. But, as we have just shown, there is no way for C to do this in equilibrium. The intuition is that to avoid war, the weak C would have to show his weakness and the tough C would have to show his strength. But they cannot do so credibly because if D were convinced that C is tough (and therefore was sure to capitulate), the weak C would have incentives to pretend he is tough and would enjoy D's capitulation to his bluff. Because the weak C has these incentives, the tough one cannot convince D that he is not lying. As we shall see, knowing something the opponent does not (private information) and having incentives to misrepresent what you know constitute one of the main explanations of why war occurs.

4.2 Pooling Equilibria

Let's consider the two remaining pure strategies for C. Suppose he plays $(1, 1)$; that is, he escalates for sure regardless of type. In this case, D's posterior belief will be:

$$x = \frac{(1)p}{(1)p + (1)(1 - p)} = p.$$

That is, the posterior belief is the same as the prior belief. This makes sense: if both types are sure to escalate, observing escalation does not tell D anything new, so her belief must remain unchanged. In this case, D's optimal behavior depends on the his prior, p. There are two cases to consider:

1. $p < 0.8$, in which case D's best response is to play r ($q = 1$). Is C's strategy $(1, 1)$ a best response to $q = 1$? It is not. Consider the tough type's expected payoffs. If C_T escalates, his expected payoff given that D will resist

and he will attack is $U_{C_T}(\alpha = 1) = -1$, which is worse than his expected payoff from not escalating at all, which equals $U_{C_T}(\alpha = 0) = 0$. Thus, the tough type would not play this strategy, and this profile cannot be a sequential equilibrium.

2. $p > 0.8$, in which case D's best response is to play $\sim r$ ($q = 0$). Is C's strategy $(1, 1)$ a best response to $q = 0$? Compare the expected utilities for the two types given that D would not resist:

$$U_{C_T}(\alpha = 1) = 10 > U_{C_T}(\alpha = 0) = 0$$
$$U_{C_W}(\beta = 1) = 10 > U_{C_W}(\beta = 0) = 0$$

Yes, this strategy is a best response. Therefore, we have found our first solution: the profile $\langle(1, 1), 0\rangle$ is a sequential equilibrium if $p > 0.8$. In words, if D believes that the chance of the challenger being tough is more than 80%, then we would expect her to back down if challenged, and therefore expect the challenger to escalate. Intuitively, since D is too pessimistic, even weak challengers can get away with escalation. Deterrence will certainly fail if the defender is believed to be pessimistic. However, the probability of war will be zero: the game will end with capitulation by the defender.

We now have a unique solution provided that $p > 80\%$.[2] We have discovered that if D is sufficiently pessimistic about the chance of her opponent being weak,

---

[2]Strictly speaking, we also need to consider $p = 0.8$. In this case D is indifferent between her two strategies because they are both best responses. Hence, D is free to mix between them with any $q \in [0, 1]$. Suppose then that D plays a mixed strategy q. Is $(1, 1)$ a best response to it? Let's compute the expected utility of the tough challenger:

$$U_{C_T}(\alpha = 1) = q(-5) + (1 - q)(10) = 10 - 15q.$$

Since in equilibrium this must be better than not escalating whose expected payoff is $U_{C_T}(\alpha = 0) = 0$, it follows that $10 - 15q > 0$ requires that $q < 2/3$. Thus, as long as the probability of resistance is less than two-thirds, escalating is optimal for the tough challenger. Let's compute the expected utility for the weak challenger:

$$U_{C_W}(\beta = 1) = q(-10) + (1 - q)(10) = 10 - 20q.$$

Since in equilibrium this must be better than not escalating whose expected payoff is $U_{C_W}(\beta = 0) = 0$, it follows that $10 - 20q > 0$ requires that $q < 1/2$. Thus, as long as the probability of resistance is less than one-half, escalating is optimal for the weak challenger. Putting these two results together, we conclude that if $q < 1/2$ (which automatically means that $q < 2/3$ as well), C's strategy $(1, 1)$ would be a best response to q. Thus, we have found other solutions. The strategy profiles $\langle(1, 1), q < 1/2\rangle$ are sequential equilibria if $p = 0.8$. There is an infinite number of these equilibria because there is an infinite number of possible $q < 1/2$ that D could pick. In these equilibria, the probability of war may not be zero. In fact, the probability of war equals exactly q, so it can be anything from $1/2$ to nothing. The multiplicity of solutions is a bit of a problem because it means that we cannot say exactly which solution would be selected. However, $p = 0.8$ is a knife-edge condition that is quite unlikely to be satisfied in practice. The likelihood that the prior beliefs

13

then the unique sequential equilibrium of the game involves deterrence failure: the challenger will escalate (even if weak), and the defender will capitulate. Observe that D capitulates even though she knows that the escalation could be a bluff (that's because she knows it could have been caused by a weak opponent). However, the risk of war is too great, so resistance is very likely to turn out a costly mistake. The defender is unwilling to take this particular bet, and capitulates instead.

Suppose now that C's strategy is $(0, 0)$; that is, C never escalates regardless of his type. In this case, D cannot update her beliefs if escalation occurs because escalation is a zero-probability event. What is D to believe then? There cannot be a sequential equilibrium in which D updates to believe that C is tough. To see this, note that if D did update this way ($x = 1$), then she would definitely back down. But if she is certain to back down, both types would prefer to escalate, so the strategy $(0, 0)$ cannot be a best response.

The only way to ensure that neither type escalates in equilibrium is that D updates to believe that the probability of a weak challenger is extremely high whenever she observes unexpected escalation. For example, if D updated to $x = 0$, then her best response would be to resist, in which case neither type of challenger would want to escalate. Although this is a sequential equilibrium, it does not seem reasonable: D is "threatening with beliefs". That is, she seems to be able to threaten C by saying "I expect you not to escalate, but if you do escalate, then I will believe that you are weak." Such beliefs are not credible because if any type would for some reason ever escalate, it would have been more likely to be the tough type, not the weak one. What to do for these equilibria is an unresolved issue in game theory. There are increasingly stronger refinements of the solution concept that eliminate unreasonable beliefs much in the same way subgame perfection eliminates unreasonable actions. Many, and sometimes all, of these weird equilibria supported by strange beliefs can be eliminated by some such refinement. However, these are beyond the scope of this course or our needs. We shall simply ignore such bizarre solutions.

Strategies, like $(1, 1)$ and $(0, 0)$, that prescribe the same actions for all types are called **pooling** because all types of C "pool" on the same behavior: they all do the same thing. Of course, if C plays a pooling strategy, D cannot infer anything new from his behavior, as we have already seen. A sequential equilibrium in which players play pooling strategies is called a **pooling equilibrium**. This game a unique pooling equilibrium if $p > 0.8$. It is reasonable to expect that deterrence will fail when the defender believes that the challenger is tough with high probability. Deterrence fails because the defender cannot credibly threaten to resist given her own beliefs. The defender's position is weakened because she thinks that her chances are not good. We should therefore expect that a lot of foreign policy will consist of

---

will be exactly equal to some particular number are vanishingly small. If p differed from 80% by even the tiniest amount, then none of these solutions will exist. Therefore, it is safe to ignore this case altogether.

bravado, swagger, and posturing where nations attempt to demonstrate that they believe they are tough and invincible. Conversely, a lot of intelligence work would be directed at estimating whether they have any basis for such posturing. You should now understand why nations care about their image. They are afraid that if they appear to believe themselves to be weak, a potential opponent will conclude that aggression will not be resisted, and will therefore proceed to challenge them.

4.3 Semi-Separating Equilibria

Up to now we only considered pure strategies for C. How about mixed ones? That is, profiles with (α, β) where α, β are some numbers between 0 and 1. As it turns out, these give us the most interesting results from our model. Again, there are several cases to consider:

1. Suppose that C plays $(\alpha, 1)$; that is, he escalates with probability $0 < \alpha < 1$ if tough, and escalates for sure if weak. We now show that no such profile can be a sequential equilibrium. The intuition is that if the weak type finds it optimal to escalate with certainty, it cannot be the case that the tough type finds it optimal to not escalate with positive probability.

   We do a proof by contradiction.[3] Suppose that $\langle (\alpha, 1), q \rangle$ is a sequential equilibrium (we don't know what q is right now). This means that escalation is a best response for C_W, which implies that:

   $$U_{C_W}(\beta = 1) > U_{C_W}(\beta = 0)$$
   $$q(-10) + (1 - q)(10) > 0$$
   $$q < \tfrac{1}{2}.$$

   That is, the fact that the weak type's strategy is optimal implies that D's equilibrium probability of resisting must be less than $\tfrac{1}{2}$. Let's compute the expected utilities of the tough type. We know that by not escalating he would get $U_{C_T}(\alpha = 0) = 0$. By escalating, he would get:

   $$U_{C_T}(\alpha = 1) = q(-1) + (1 - q)(10) = 10 - 11q,$$

---

[3]A proof by contradiction works as follows. Suppose we want to prove that some statement is false. We assume that it is true and then demonstrate that it being true implies something that is contradictory. We can therefore conclude that the statement cannot be true; i.e. that it is false, which is what we wanted to show. Here's an example. We know that I only teach National Security Strategy (NSS) on Mondays, Wednesdays, and Fridays, each day at 11:00a. Consider the statement "If it is 11:00a, then I am teaching NSS." We want to prove that this statement is false. Assume that it is true, seeking a contradiction. Since it is true, it implies that I am teaching NSS on Sundays as well (because there is no reference to the day of the week in the statement). But we know that I only teach MWF, which implies that I do not teach NSS on Sunday. Hence, we arrive at a contradiction. We conclude, that the statement "If it is 11:00a, then I am teaching NSS" is false.

and because $q < \frac{1}{2}$, this means that:

$$U_{C_T}(\alpha = 1) = 10 - 11q > 10 - 11(\frac{1}{2}) = 4.5 > 0 = U_{C_T}(\alpha = 0).$$

In other words, if $q < \frac{1}{2}$ (which must be the case since the weak type is escalating), the tough type can always get a higher expected payoff from escalating than from not escalating himself. This means, that it cannot be optimal to play a strategy that puts positive probability on not escalating. Thus, there can be no sequential equilibrium where C plays a strategy $(\alpha, 1)$.

2. Suppose now that C plays $(\alpha, 0)$; that is, he escalates with probability α if tough, and never escalates if weak. Because the weak type never escalates, D's posterior probability is $x = 1$, and so her best response is to capitulate. But if D is going to capitulate for sure, then the weak type would strictly prefer to escalate, $\beta = 1$. Thus, there can be no sequential equilibrium where C plays a strategy $(\alpha, 0)$.

3. Suppose now that C plays $(0, \beta)$; that is, he never escalates if tough, and escalates with probability $\beta > 0$ if weak. Because the tough one never escalates, D's posterior belief will be $x = 0$, and so her best response would be to resist. But if D is expected to resist for sure, then the weak type would strictly prefer not to challenge her at all, and so $\beta = 0$. Thus, there can be no sequential equilibrium where C plays a strategy $(0, \beta)$.

4. Suppose now that C plays (α, β); that is he escalates with probability α if tough, and with probability β if weak. Suppose that D resists with probability q (we don't know what it is for now). Since the weak type is willing to play a mixed strategy, he must be indifferent between his two actions. That is,

$$U_{C_W}(\beta = 1) = U_{C_W}(\beta = 0)$$
$$q(-10_+(1-q)(10) = 0$$
$$q = \frac{1}{2}.$$

Thus, if the weak type is willing to randomize, it must be the case that $q = \frac{1}{2}$ (or else one of the pure actions would yield a strictly higher payoff, and C_W would choose to play it with certainty). However, if $q = \frac{1}{2}$, we have:

$$U_{C_T}(\alpha = 1) = q(-1) + (1-q)(10) = 4.5 > 0 = U_{C_T}(\alpha = 0).$$

That is, the tough type's expected utility from escalating is strictly better than his expected utility from not escalating. This means that the tough type would not be willing to play a mixed strategy which puts a positive probability on not escalating. Therefore, $\alpha < 1$ cannot be an optimal strategy. Hence, there is no sequential equilibrium where C plays a strategy (α, β).

5. This leaves one final possibility to consider. Suppose now that player C plays a strategy $(1, \beta)$; that is, he escalates for sure if tough, but escalates only with probability β if weak. Let q denote the mixed strategy for D. As we have already seen, since the weak type is willing to randomize, his expected payoff from escalating must equal his expected payoff from not escalating. That is, it must be the case that $q = 1/2$. We have already seen that if $q = 1/2$, then the tough type strictly prefers to escalate, so the strategy $\alpha = 1$ is optimal, as specified here.

Since D herself must be willing to randomize to play $q = 1/2$, it follows that her expected payoff from resisting equals her expected payoff from not resisting. If this were not the case, she would always choose the strategy that yielded the higher payoff. Thus, it must be the case that $U_D(q = 1) = U_D(q = 0)$, or:

$$U_D(q = 1) = U_D(q = 0)$$
$$x(-15) + (1 - x)(10) = x(-10) + (1 - x)(-10)$$
$$-25x = -20$$
$$x = 0.8.$$

Noting that D's posterior belief will be $x = \frac{p(1)}{p(1) + (1-p)\beta}$, we now have:

$$\frac{p}{p + (1 - p)\beta} = 0.8.$$

Solving this for β yields:
$$\beta^* = \frac{p}{4(1 - p)}.$$

In other words, if the weak challenger chooses to escalate with probability β^*, he will induce in D the belief $x = 0.8$, which will make her indifferent between her two strategies, which in turn rationalizes her randomization. This is an instance of how a rational player can manipulate the beliefs of a rational opponent. Clearly, there is not much latitude in doing so, as we should have expected. It should not be too easy to get a rational player to believe whatever one wants.

We now have the exact mixing probability for the weak type (as a function of the prior belief p) that would yield the necessary posterior belief for D, which would in turn make C's strategy optimal. Here's how this works. Given D's prior belief p, the weak challenger will choose a mixing probability β^*, which will ensure that the defender will be indifferent between resisting and not resisting, and will be willing to mix between them. In particular, it could be optimal to do so with probability $q = 1/2$, which in turn renders the weak

challenger's strategy optimal. We already know that if D mixes like that, the tough challenger will always escalate.

Note now that β^* has to be a valid probability. It is positive, so we only need to ensure that $\beta^* < 1$. Solving this inequality for p yields the condition that $p < 0.8$. Thus, manipulating D's belief in this way is only possible if her prior belief assigns less than 80% chance to C being tough. The reason for that is simple: to get D to capitulate, she must believe that it is quite likely that her opponent is tough (in this case, the probability must be at least 80%). With a strategy according to which the tough challenger is more likely to escalate than the weak one, the posterior belief will always exceed the prior. That is, after escalation D will become more pessimistic. If the prior is already above 80%, then she is already pessimistic enough to begin with and there is no need to manipulate her belief: the challenger escalates regardless of type and reaps the benefits. Only when D starts out relatively optimistic that C must manipulate her beliefs.

Thus, the profile $\left\langle \left(1, \frac{p}{4(1-p)}\right), \frac{1}{2} \right\rangle$ is a sequential equilibrium if $p < 0.8$. Note that this equilibrium exists only for values of the prior probability p (we can always compute the necessary β from it) that are less than 80%, unlike the pooling equilibrium we found before which only existed for values that exceed 80%. We shall analyze the substantive features of this equilibrium in the next section.

Before we analyze the equilibrium we found, let's give it a name. A strategy in which one type plays some action with certainty and another type plays that action with positive probability is called **semi-separating**, or "partially separating." This is because the two types only partially separate themselves by their behavior. If D observes escalation, she can update the probability of her opponent being tough because the tough type is more likely to have escalated, but D still cannot be absolutely certain. Some information gets transmitted, but not enough to ensure D of the type of opponent she is facing. (If D does not observe escalation, then she can conclude that the opponent is weak because the tough types always escalate, so the status quo is only kept by the weak type with positive probability.) A sequential equilibrium in which players play semi-separating strategies is called a **semi-separating** (or "hybrid") equilibrium. This game has exactly one such equilibrium.

5   Substantive Implications

What does all this analysis give us in terms of substantive ideas we can use when thinking about real crises? We have already learned two things from the pooling equilibrium and the nonexistence of separating equilibria:

- We cannot reasonably expect states to reveal their private information in a crisis because weak ones always have incentives to pretend they are strong. No amount of diplomacy and communication can alter this basic fact.

- Deterrence can fail if the defender appears pessimistic of her chances, which means that states would tend to publicly brag about their strengths and exaggerate the weaknesses of their adversaries.

- Because states cannot credibly communicate capabilities or intentions, they will spend a lot of resources on uncovering information about their opponents (and preventing their opponents from learning stuff about themselves). But even after a state finds evidence that its opponent is weak (e.g. satellite photos show few missiles), the opponent can dismiss public statements to that effect by arguing that the state would have said something like this anyway because it is in its interest to pretend that its opponents are weak.

Note now that our solutions are very general because they cover the entire range of initial beliefs p that the defender might have. If $p > 0.8$, then the model predicts the pooling equilibrium in which the challenger escalates regardless of type and the defender backs down. The equilibrium outcome will be surrender by the defender. This is the certain deterrence failure equilibrium.

If $p < 0.8$, on the other hand, the model predicts the semi-separating equilibrium, in which tough challengers always escalate, but weak ones only do so sometimes. The defender sometimes resists and sometimes does not. This is an equilibrium that may involve deterrence failure and success, and, more importantly, it may involve war with positive probability.

Let's now see what the semi-separating equilibrium tells us. First of all, what is the equilibrium outcome? The probability that C will escalate, as we have already seen, is given by the formula:

$$\Pr(e) = \alpha p + \beta(1 - p) = 1.25p,$$

where $p < 0.8$ (again, recall that $p < 0.8$ is a necessary condition for this equilibrium to exist). The probability of non-escalation (deterrence success) is simply the probability that the challenger is weak and does not escalate (the tough one always does):

$$\Pr(\sim e) = (1 - \beta)(1 - p) = 1 - 1.25p.$$

The probability of war is the probability that the tough one escalates and the defender resists:

$$\Pr(\text{War}) = \Pr(C_T) \times \Pr(e|C_T) \times \Pr(r) = p \times (1) \times 0.5 = 0.5p.$$

The probability that the defender capitulates is:

$$\Pr(\text{Cap}_D) = \Pr(C_T) \times \Pr(e|C_T) \times \Pr(\sim r) + \Pr(C_W) \times \Pr(e|C_W) \times \Pr(\sim r) = 0.625p.$$

Finally, the probability that the challenger capitulates (compellence) is:

$$\Pr(\mathrm{Cap}_C) = \Pr(C_W) \times \Pr(e|C_W) \times \Pr(r) = 0.125p.$$

Let's do a quick check of our calculations. Deterrence can either succeed or fail. That is, either escalation or non-escalation should occur: $\Pr(e) + \Pr(\sim e) = 1.25p + 1 - 1.25p = 1$, so we're fine. The possible outcomes are status quo (non-escalation), capitulation by defender, capitulation by challenger (compellence), and war: $\Pr(\sim e) + \Pr(\mathrm{Cap}_D) + \Pr(\mathrm{Cap}_C) + \Pr(\mathrm{War}) = 1$, so we're fine.

   We now turn to implications. First, note that the probability of war is strictly positive and increasing in the defender's pessimism up to a point (when $p = 0.8$). That is, if the defender believes that her opponent is unlikely to be tough (p is low), the chance of war is also low because the defender becomes likely to resist, and therefore unlikely to be challenged. As the defender's pessimism increases, the danger of war begins to loom larger and larger. It is highest right before the defender crosses the threshold of believing with 80% chance that his opponent is tough. At this extremely dangerous point, the probability of war is close to 40%.

   Note some rather telling dynamics here. As p increases (defender's pessimism goes up), the probability that she will capitulate also increases. However, the probability that the challenger will capitulate increases as well. This is because when the probability of non-resistance by defender goes up, more weak challengers will test their luck by escalating. But since the defender sometimes does resist, this means that more of them will end up capitulating themselves. Unfortunately, this means that she will also sometimes resist escalation by a tough opponent causing war.

   This is very interesting: the weaker the defender believes herself to be, the more vulnerable she will appear to challengers, who may miscalculate and escalate. But because the defender still believes that the challenger might be weak, she may respond by resisting, which would compel a weak one to capitulate. Unfortunately, it would also cause war if the challenger is genuinely tough. The weaker the defender appears, the more likely are weak challengers to try their luck, which implies that the defender is herself more likely to resist because of the higher chance of the challenger being weak. But this, paradoxically, increases also the chance of war.

   Once the threshold of $p = 0.8$ is crossed, the defender becomes hopelessly pessimistic, and never resists. Suddenly the chance of war drops precipitously to zero even though all types of challengers now escalate. Deterrence always fails but no war will occur as a consequence. Thus, it is incorrect the state that the probability of war always increases with pessimism. Rather, it increases up to a point, and then abruptly goes down. It is also incorrect to state that optimism generally causes war. As we have seen, the probability of war can drastically jump at the point of the equilibrium switch from pooling to semi-separating. However, from this point on as optimism increases (p goes down), the probability of war will decrease.

   This escalation model is a **signaling game**, because the **informed player** (the challenger) gets to move first and reveal something about his privately known type.

Of course, as the pooling equilibrium demonstrates, sometimes no information will get transmitted. Further, as the lack of separating equilibria shows, it is not possible for all information to be revealed. However, there are many cases where the defender can learn something about her opponent, as in the semi-separating equilibrium.

The tough opponent tries to signal that he is tough by escalating in the semi-separating equilibrium. Indeed, the posterior belief x is always greater than the prior belief p regardless of β. Because tough types escalate more than weak ones, escalation is an imperfect signal that the challenger is tough. Surprisingly, this does not cause D to alter her behavior: Her equilibrium strategy is to resist and capitulate with equal probability. This makes sense: If she responded by capitulating more often, then she would encourage more weak types to try escalation in the hopes that they will get lucky. So the defender balances the risk of war with the desire to deter weak opponents from escalating. As you can already see, deterrence and compellence are balancing acts that are difficult and counter-intuitive (without the model).

It is worth emphasizing one other central conclusion. War is really bad for both the defender and the challenger regardless of the challenger's type. If the players had a choice between living with the status quo and fighting, *both would prefer to live with the status quo*. In other words, both players are "peace-loving" because neither likes war for its own sake. One often hears the argument that "only if everyone loved peace, then we would not have war" and its extension to "everyone that starts a war must be a war-monger."

Our model demonstrates that this claim and entire line of reasoning is incorrect. Both our players are peace-loving and yet the probability of war is strictly greater than zero as long as the defender is not entirely pessimistic (that is, as long as $p < 0.8$). Thus, war can occur despite both players hating it. The problem here is not whether one likes peace or not, but whether one is prepared to risk war to prevent an opponent from taking advantage of one's love of peace. As the Romans said, if you want peace, prepare for war.

We shall see this trade-off between the risk of war and the gain from the threat to unleash it quite often. It even has a formal name in the theory of the use of force as the **risk-return trade-off**. It captures the idea that one would pursue policies that balance the risk of disaster with the gains from threatening it. That is, one would press one's advantage to the extent that doing so would increase one's gains, which always comes at the expense of a increased risk of disaster. At some point, you would forego additional gains because the risk becomes intolerable. This is the trade-off. But of course, when you pursue such a strategy, you do actually run the risk of everything ending in disaster.

We conclude that **it is entirely rational for peace-loving nations to end in a war with each other**. Destruction, even when hated by everyone, might occur and we do not need irrationality or evil to explain it. It is a natural consequence of

players trying to obtain good outcomes in the international arena. The reason that war does not occur all that often should also be intuitive: it is precisely because of its costliness that even a small risk of disaster is sufficient to deter most players from attempting to extract more from their opponents.

The presence of incomplete information is a necessary condition for rationalizing war in our model. Thus, we can conclude that **private information and incentives to misrepresent it constitute a major cause of war**. When you read the article by James Fearon, you will note that this is one of his two central claims. The article was written in 1995 and you should carefully note that its author goes through a long list of causes of war that people have proposed and finds problems with all explanations.

6   Coming Up Next...

Note that the idea of credibility pervades the entire discussion because it is a fundamental feature of the solution concept we employed. The main concerns are: (a) how can the tough challenger credibly reveal its type? (b) how can the defender credibly commit to resisting, thereby enhancing the prospect of deterrence success? (c) how can a challenger credibly commit to compelling the defender to back down? All these turn out to be inter-connected issues and we cannot consider one of them in separation from the others. The solution to the model showed why this is the case. It also uncovered some rather intriguing dynamics resulting from uncertainty. Next time we shall look at how one can, at least in part, overcome some of the problems caused by informational asymmetries.