Sequential Moves and Credibility

Every non-zero-sum conflictual situation is an occasion for bargaining. Players will try to influence each other's expectations because the actions they take depend on what they know about each other and how they expect the opponent to respond. Players commit to certain actions in the future (either threats or promises) in an attempt to induce the opponent to change his expectations. For the opponent to alter his expectations, he must (a) understand the commitment, and (b) believe it. Therefore, players must be able to **communicate** their commitments in a way that will ensure that the opponent understands them, and they must make sure that their commitments are **credible** so that their opponent will believe them.

1 Massive Retaliation

In the 1950s, the United States enjoyed an enormous advantage in nuclear forces over the Soviet Union and it relied on a strategy called Massive Retaliation. According to this strategy, a Soviet military challenge in Europe or anywhere else would be met with a nuclear strike on USSR. In the 1960s, the Soviet Union acquired enough capability to retaliate with a devastating nuclear strike even after absorbing a first strike by the United States. This is called *secure second-strike capability*. The U.S. also had it.

Figure 1 shows the payoff matrix for this situation. C stands for "Challenge" and $\neg C$ stands for "Don't Challenge". A stands for "Attack with Nuclears" and $\neg A$ stands for "Don't Attack". If the USSR does not challenge the status quo, then both players receive a payoff of zero. If, however, the USSR challenges successfully, it gains 5 while the U.S. loses 5. On the other hand, a nuclear war is devastating for both sides (because they both have secure second-strike capability).

		U.S. A	U.S. $\neg A$
USSR C		$(-10,-10)$ nuclear war	$(5,-5)$ successful challenge
USSR $\neg C$		$(0,0)$ threat, no challenge	$(0,0)$ no threat, no challenge

Figure 1: Massive Retaliation in Secure Second-Strike Capabilities.

This game has two Nash equilibria in pure strategies: $(A, \neg C)$ and $(\neg A, C)$:

- $(\neg C, A)$: the U.S. threatens to launch a massive strike in response to a challenge and the USSR does not challenge.

  To see that this is an equilibrium, consider the best responses for both players. The U.S. best response to $\neg C$ is either A or $\neg A$ because they both yield the same payoff of 0. However, since USSR's best response to $\neg A$ is C, not $\neg C$, the strategy profile $(\neg C, \neg A)$ is not an equilibrium. On the other hand, USSR's best response to A is $\neg C$, and so $(\neg C, A)$ is a profile of mutual best responses.

  Given the strategy of the other player no player has an incentive to deviate from its own strategy. The Eisenhower administration maintained that Massive Retaliation would work exactly like this. No challenges occur in this equilibrium, and since the USSR never challenged the status quo militarily, some think that Massive Retaliation worked as advertised.

- $(C, \neg A)$: the U.S. does not threaten to launch a strike and the USSR challenges the status quo successfully.

  To see that this is an equilibrium, consider the American best responses first. The U.S. best response to C is $\neg A$, and since the Russian best response to $\neg A$ is C (as we have already seen), it follows that the strategies are mutually best responses.

  This is an equilibrium in which the U.S. acquiesces to a military challenge to the status quo by the Soviet Union. Critics of Massive Retaliation (many European allies were among them) maintained that this would be the result of the policy, not the successful deterrence as the supporters claimed based on reasoning derived from the above equilibrium.

Who was right? The Eisenhower administration or the opponents of Massive Retaliation?

2 Massive Retaliation with Sequential Moves

The purpose of the United States was to deter the Soviets from challenging the status quo. That is, Massive Retaliation was meant to serve as a threat that would cause the Russians to abandon any adventurous ideas for expansion around the globe. The way deterrent threats are supposed to work is quite simple: once the U.S. has issued the threat, the Soviet Union gets to decide whether to challenge or not, and if it challenges, the U.S. must decide whether to follow through on the threat.

This is a situation in which moves are made in sequence. Assume that the U.S. has already made the deterrent threat of massive retaliation. Figure 2 depicts the sequence of moves beginning with the decision of the Soviets.

The game still has the two Nash equilibria. We can easily see the first one, $(C, \neg A)$ by **looking forward and reasoning backward**, a process formally known as **backward induction**.

To determine the players' choices, we begin at the end of the game with the player who gets to move last. In our example, this would be the U.S. having to respond to a Soviet challenge. We look at the payoffs: if the U.S. initiates massive retaliation (which leads to

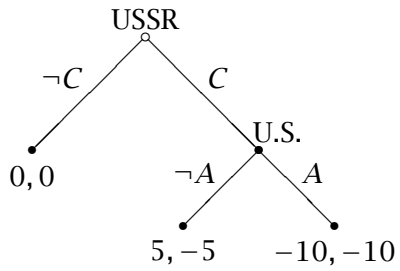

Figure 2: Massive Retaliation in Extensive Form.

nuclear war), its payoff is -10. If, on the other hand, it does not follow through on its threat and lets the Soviets get the payoffs of a successful challenge, its payoff is -5. Therefore, *if the Soviets ever choose to challenge the status quo, the United States will never initiate a nuclear strike.*

Given that the United States chooses to back down, the Soviets' payoff from challenging is 5, while the payoff from not challenging is only 0. Therefore, the Soviets will challenge and the U.S. will respond by acquiescing. The outcome $(C, \neg A)$ is called a **subgame perfect equilibrium** because the strategies of the players induce an equilibrium in every possible subgame. (A subgame is just any part of the game tree that begins with a move by some player.)

The subgame perfect equilibrium is always a Nash equilibrium as well, but the converse is not true. That is, not all Nash equilibria are subgame perfect. In particular, the other Nash equilibrium $(\neg C, A)$ is **not** subgame perfect because the strategy it specifies for the U.S. is not optimal in the subgame that begins with the American decision (that is, the subgame that begins after a Soviet challenge).

The problem with this Nash equilibrium is that it is supported by a **incredible** threat by the U.S., and so it is unreasonable. This bears repeating: the $(\neg C, A)$ strategy profile *is a Nash equilibrium:* if the Soviets play $\neg C$, then the U.S. decision node *is never reached* and so regardless of what the U.S. chooses to do at its decision node, it would still receive zero because the Soviets never challenge the status quo. Since every strategy is a best-response, the U.S. has no incentive to deviate from A if the USSR plays $\neg C$, and since the U.S. plays A, the Soviet best response is $\neg C$.

The problem with Nash equilibrium is that it only examines that optimality of decision along the **equilibrium path**. That is, it only ensures that strategies are best responses to each other for the choices they specify. The solution concept ignores actions **off the equilibrium path**, that is, it ignores situations that will never arise if players follow their equilibrium strategies.

For example, if players follow their equilibrium strategies in $(\neg C, A)$, then the equilibrium path is just the action $\neg C$ by the Soviets with the status quo outcome. The U.S. decision is off the equilibrium path because it is never reached when players follow their strategies. Nash equilibrium simply ignores what happens in these situations.

However, as we can clearly see in this example, this is a problem. The problem is that the Soviet strategy $\neg C$ is only optimal if they believe the Americans are going to carry out the threat A if they are faced with a challenge. But the Americans never will, and so it is

unreasonable to suppose that the Russians will then stick to $\neg C$!

3 Subgame Perfect Equilibrium

Nash equilibrium only follows the equilibrium path of play and never gets to that American decision, and so it never ensures its optimality. However, the process of backward induction guarantees that we examine all possible situations, and so it eliminates this Nash equilibrium as unreasonable, leaving us the only one where all commitments are credible.

In order to determine optimal play in any situation, we need to examine ALL possible other situations that might arise if we (or our opponent) take alternative actions. This includes situations that might not arise if you follow your own strategy!

Let's see why this is the case. Figure 3 modifies the Massive Retaliation game by adding an initial move by the United States, which can now choose to issue the threat publicly or not. In either case, the two players then play the Massive Retaliation game from before. The difference is that if the U.S. publicly issues a threat and then does not follow through on it, it loses face and pays a reputational cost of 2. (We shall list the payoffs as before, with the first one for the USSR and the second one for the U.S. Usually, the payoffs are listed according to the order of moves, with the first number being the payoff to the player who gets to move first, the second, for the one who moves next, and so on).

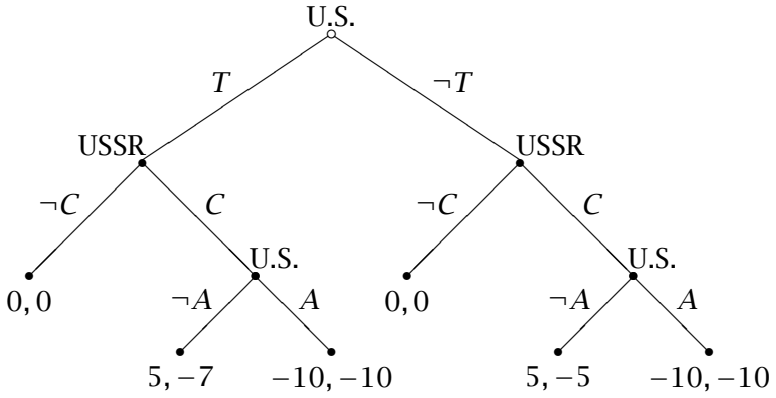

Figure 3: The Massive Retaliation Threat.

Let's find the subgame perfect equilibrium by backward induction. Start with the U.S. decision at the left penultimate node (the one following the path of play T, C). Here, the U.S. would prefer $\neg A$ (payoff of -7) to A (payoff of -10). At the right node, the U.S. prefers $\neg A$ (payoff of -5) to A (payoff of -10) also. In other words, whenever the U.S. has to respond to a Soviet challenge, it will always back down instead of starting a general nuclear war.

Given this strategy, what are the Soviets to do? At both decision nodes the USSR strictly prefers to challenge because it gets 5 instead of 0 if it chooses $\neg C$. So, given that the Americans are expected to acquiesce, the Soviets will always challenge.

Let's now finally move up the tree to the initial decision by the U.S. whether to issue a threat or not. If it issues the threat (T), it knows that the Russians will play C, to which

the U.S. would have to respond with $\neg A$, yielding the path of play $T, C, \neg A$, which gives the U.S. a payoff of -7 because of lost reputation in addition to lost interests. If, on the other hand, the U.S. chooses $\neg T$, the Russians will respond with C, to which the Americans would respond with $\neg A$. The path of play is $\neg T, C, \neg A$ and the payoff is -5. Still lost interests but without the reputational loss. Given these strategies, the U.S. strictly prefers to stay quiet and not issue a useless threat at the outset.

The subgame perfect equilibrium is therefore the following pair of strategies:

- U.S.: $\neg T$ at first node, and then $\neg A$ if the Soviets challenge without a threat, and $\neg A$ of the Soviets challenge after a threat; this can be compactly written as $(\neg T \neg A \neg A)$.
- USSR: C if the Americans threaten and C if they don't, compactly written as (CC).

Note how the strategies specify what the players must do in **all possible contingencies, even ones that do not arise if they follow their equilibrium strategies.** For example, we specified the action of the U.S. given that the USSR has challenged after an American threat even though if the U.S. follows its equilibrium strategy it will never issue the threat in the first place.

You should, however, very intuitively see the reason for this. When you were doing the backward induction, you had to consider all possible subgames (all possible situations) before you could determine the optimal move of the U.S. at its initial node. You had to do this because what the U.S. does at the first node depends on what it thinks the Soviets would do, which in turn depends on what the Soviets think the Americans would do next. *The the first move by the U.S. is only optimal given the optimal behavior of the Soviets, which in turn is determined by the optimal behavior of the U.S. at the end of the tree.*

That is why we write the strategy profile in full, listing all optimal actions at every possible place where an actor has to make a decision.

The general point is that to determine the optimality of an action, we have to examine all possible contingencies, even the ones that will not arise if the action is taken. Again, without considering all contingencies, we cannot form expectations about the behavior of our opponent, and therefore will not be able to make optimal decisions in the first place. Very intuitive stuff.[1]

Remember this: Nash equilibria may rely on incredible commitments, and so may produce unreasonable predictions in situations with sequential moves. Subgame perfection imposes the requirement that strategies are optimal in all possible situations in the game.

[1] Back to Massive Retaliation, you can now see clearly that regardless of what benefits the administration claimed for its strategy, the critics were correct in their analysis that the strategy would not work. The reason it would not work was quite simple: the threat of a general nuclear war in retaliation for conventional military incursion was simply not credible. And because it was not credible, there was no reason to expect the Russians to believe it. And since the Russians would not believe it, the U.S. had to find other ways of deterring their possible aggression. The Americans, by the way, did design a host of other strategies, of which you can learn if you take my National Security Strategy class next year.

An interesting consequence of this analysis is that since the Russians did not believe the deterrent threat, their quiescence during the long decades of the Cold War must be attributed not to fear of the Americans but to something else. Given some recently declassified Soviet materials, it appears that they simply did not have many of the aggressive world conquest aspirations the Americans thought they did. That is, it may have been the case that the Soviets did not invade Western Europe simply because they did not want it.

This eliminates these undesirable equilibria. You can find the subgame perfect equilibria by backward inductions. All subgame perfect equilibria are also Nash equilibria, but not all Nash equilibria are subgame perfect. **Subgame perfection eliminates unreasonable Nash equilibria that are based on inherently incredible threats.**

4 North Korea, Nuclear Weapons, and Bribes

North Korea decides whether to build nuclear facilities for enrichment of uranium. If it does not, the status quo prevails and both it and the U.S. get payoffs of zero. If it builds, North Korea pays $10 million for the cost of the plant. The United States can offer it a bribe worth π million dollars to dismantle the plant (assume that you can only choose π in $1 million dollar increments). If the North Koreans accept, the payment is made and the plant is dismantled. If they reject, the United States can initiate an air strike to destroy the plant, which causes an international incident that costs both sides $30 million. If the U.S. acquiesces, the North Koreans enjoy $20 million worth of plant operation net of $10 million in costs, and the U.S. loses $25 million in having to build up its defenses. Assume that if the North Koreans are indifferent between dismantling the plan and keeping it, they always keep it.

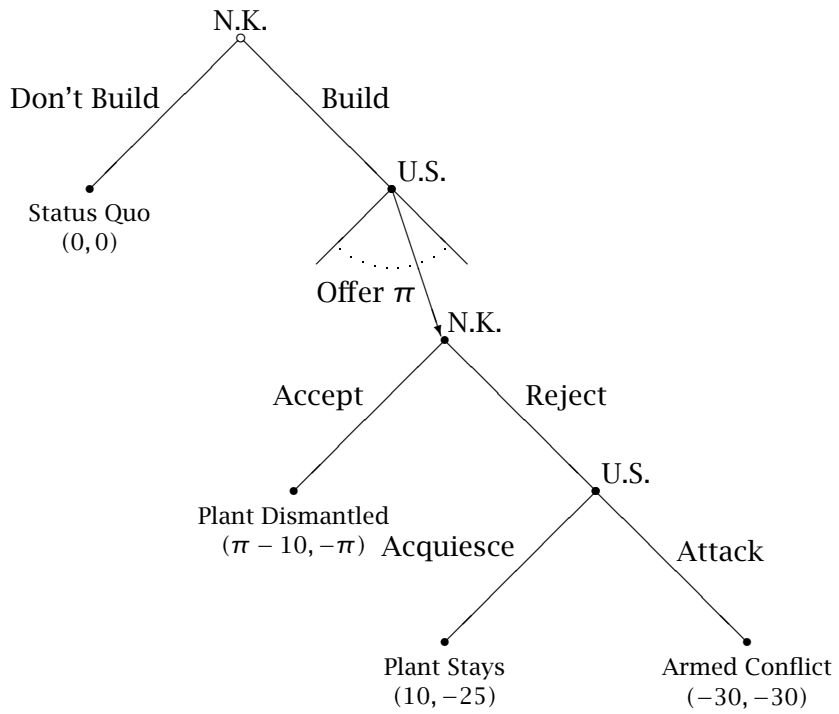

Figure 4: Bribing the North Koreans, I.

Let's now analyze the situation. In particular, we want to know

- Can the U.S. bribe the North Koreans to get them to dismantle the plant?

- If it can, how much should it cost?
- Can the U.S. adopt a strategy that will induce the North Koreans not the build the plant in the first place?

We apply backward induction. Start with the last move in the game, which is the U.S. decision to attack. Since attacking yields a payoff of -30 and acquiescing yields a payoff of -25, the U.S. would acquiesce. This means that the North Koreans know that if they reject the U.S. bribe, they will certainly get a payoff of 10 (because the Americans will not attack).

If the U.S. is to get them to dismantle the plant, its bribe must be such that the North Koreans would prefer to do it. That is, we must find a π such that $\pi - 10 > 10$, which means $\pi > 20$. In other words, the U.S. must offer more than \$20 million if the North Koreans are to agree to dismantle the plant.

But how much should the U.S. offer? Anything over \$20 gets the North Koreans to agree to a deal, which leaves the U.S. with a payoff of $-\pi$. If this is to be an optimal decision, the U.S. must be doing better with an agreement to dismantle the plant than with a rejection of its offer that yields it a payoff of -25. This means that the U.S. will never offer more than \$25 million because doing so makes it strictly worse off than offering something the Koreans would reject. We conclude that the U.S. must offer something between \$20 million and \$25 million.

Of course, the U.S. is interested in making the least costly bribe that would succeed because in case of success it has to pay it. Since we assumed that only \$1 million increments are allowed, the smallest possible bribe that would satisfy the North Koreans is \$21 million, which nets them a payoff of \$11 million if they accept versus the \$10 million if they reject. So they accept. The payoff of the U.S. is -21, which is strictly better than -25 it would have gotten in case of rejection. Offering \$21 is also optimal because any larger offer, while still accepted by the Koreans, would leave the U.S. with a strictly worse payoff.

Given that North Korea can expect a deal worth \$11 million if it builds the plant versus a payoff of 0 if it keeps the status quo, it definitely chooses to build the plant. This finishes the analysis using backward induction. Let's now put all these conclusions together to specify the equilibrium strategies and then, using the strategies, see what the subgame perfect equilibrium **outcome** of the game will be.

The subgame perfect equilibrium strategies are

- North Korea: build the plant, accept offers larger than \$20 million and reject others.
- United States: offer \$21 million and do not attack if the Koreans reject.

The equilibrium outcome is: the North Koreans build the plant, get bribed \$21 million by the U.S., and dismantle it.

This answers the first two questions we set out to answer. What about the third? Is there a strategy the U.S. can adopt that would make the Koreans not build the plant in the first place?

Suppose the U.S. could commit to attacking at its last node. In that case, it can offer $\pi = 0$ as a bribe, but since the Koreans know that if they reject, they will get -30, they would accept and dismantle the plant, getting -10. Given that, they will never build the

plant in the first place and get a payoff of 0. If only the U.S. could make its threat credible, the North Koreans would never challenge the status quo. But that, of course, is precisely the problem. This U.S. threat is *not credible*, and so the North Koreans would never believe it. And because they would never believe it, they can exploit the U.S. and force it to pay for dismantling of their nuclear plant.

Of course, we simplified the situation here by assuming that the North Koreans will actually dismantle the plant after accepting the bribe. Although we examined the credibility of the U.S. threat to attack if an offer is rejected, we did not examine the credibility of the promise that the Koreans will follow through on their agreement. You will analyze this case in your homework.

5 Summary

- we must study **sequential moves** in order to analyze how **credible** the commitments of the players are
- we use **extensive form games** to model situations where the players move in sequence
- Nash equilibria may rely on incredible threats and so may be unreasonable
- **subgame perfection** ensures that all threats are credible and eliminates the unreasonable Nash equilibria
- we find the subgame perfect equilibria by **backward induction**, which is a process of looking forward and reasoning backward
- we must take care to examine the optimality of the strategies everywhere along the game tree, even at places that will not be reached if the strategies are followed; unless players form expectations about all possible contingencies, they cannot judge the optimality of their actions
- we shall use subgame perfection to study the **credibility of commitments** (both of threats and promises)