U.S. Foreign Policy:
Prisoner's Dilemma vs. Stag Hunt

Branislav L. Slantchev
Department of Political Science, University of California, San Diego

Last updated: January 20, 2016

Among the issues we will have occasion to discuss is *trust*. More specifically, when can actors trust each other to "do the right thing", with the "right thing" being defined as behavior that produces socially improving outcomes. It will be easier to illustrate what I mean with two games that are supposed to model competitive arming, or an *arms race*. Suppose it is determined that a new technology has just emerged and that it allows both us and our enemy to produce a super weapon that can guarantee winning a confrontation against an opponent who does not have it. The confrontation is very important. If both have the weapon, the effects cancel each other out. It takes a year to construct the weapon, but once built, it becomes immediately useful. The weapon is quite costly and each nation must shift resources from consumer goods to the military sector, which is politically unattractive. Should we build the weapon or not?

The interesting thing in this setup is that we have assumed that the arms race is useless: when both actors have the weapon, neither gains over the status quo and both pay the costs of building the weapon. Thus, both actors must strictly prefer the status quo and arms races should never occur. We shall now attempt to rationalize this seemingly baffling outcome: an expensive arms race that gives neither side the advantage. (And we're not going to rely on the actors being irrational, or stupid, or mistaken.)

1 Prisoner's Dilemma: Strictly Dominant Individual Incentives

We have already simplified the situation drastically in this description, so it is easy to model. Consider first the now-familiar Prisoner's Dilemma, which is here reproduced so we can compare it with the second model more easily. Recall that this game involves two actors, A and B, who must decide whether they want to cooperate with each other or not. Here, one can cooperate, C, by not building the weapon, or defect, D, by building it. There are four possible outcomes: $\langle D, D\rangle$ (both build weapons; an arms race), $\langle C, D\rangle$ (only B builds the weapon; defeat for A and victory for B), $\langle D, C\rangle$ (only A builds the weapon; victory for A and defeat for B), and $\langle C, C\rangle$ (neither builds the weapon; status quo).

We now need to decide on the preferences the actors have over these four outcomes. Since the disputed issue is assumed to be very important and the unilateral possession of the weapon guarantees that it will be resolved in the possessor's favor, each actor most prefers to be the one who has the weapon. Conversely, the worst possible outcome is to be forced to concede the issue because your opponent has the weapon but you do not. In other words, each actor prefers victory to defeat.

The other two outcomes are intermediate, and it is reasonable to assume that both actors prefer the status quo to an arms race. The reason is that under our assumptions, when both build the weapons, their military effects cancel out, so the political status quo remains except that now both have paid the cost of arming. Under this assumption, it certainly makes sense to assume that each would rather have the status quo for free than having to pay for it.

Overall, we have assumed that the preference ordering for each actor is as follows:

$$\text{Victory} \succ \text{Status Quo} \succ \text{Arms Race} \succ \text{Defeat} \qquad \text{(Prisoner's Dilemma Preferences)}$$

We can now easily visualize the possible outcomes by making them explicitly determined by the strategies in tabular form: If you look at the preference orderings, you will see that

		Player B: C	Player B: D
Player A	C	Status Quo	Defeat for A, Victory for B
	D	Victory for A, Defeat for B	Arms Race

Figure 1: Cooperation and Defection Game.

each player's most preferred outcome is the other player's least preferred one. You might reasonably conclude that neither of these outcomes would be sustainable because the player who is supposed to cooperate unilaterally would instead build the weapon as well. Since the status quo is the second-best outcome for both players, you might then conclude that this should be the outcome produced by reasonable play. Unfortunately, this will not be the case: if a player believes that his opponent will choose to cooperate, then he is strictly better off not cooperating. In fact, not cooperating is the *strictly dominant strategy* in this scenario: it is always the best option for each player regardless of what the other player does. To see this, look at Figure 1 and consider what A is supposed to do. If he thinks that B will choose C (not build the weapon), then cooperation results in the Status Quo while defection results in outright Victory. Naturally, A will defect. If, instead, he thinks that B will choose D (build the weapon), the cooperation results in Defeat whereas defection results in an arms race. Naturally, A will defect. Thus, A strictly prefers to defect irrespective of what he thinks B will do. Thus, the only strategy that we (and B) should expect him to play is D. The same reasoning applies to B, so we (and A) can only reasonable expect her to play D too. This means that the only rationalizable outcome is $\langle D, D\rangle$, the arms race.

Pause for a minute to think what this means. We have a social situation in which both players agree that cooperating with each other is the second-best choice for both of them. Unfortunately, pursuing their individually rational strategies makes both players worse off. Rationality (at least in this sense) condemns the actors to their next-to-last preferred outcome. In this instance, they will engage in a costly arms race that will make both of them worse off relative to the status quo. They do not do this because they were stupid, irrational, or mistaken. They do this because their incentives in this situation are not aligned properly to support mutual cooperation.

One possible way to rationalize the baffling arms race is by noting that the individual incentives to take advantage of the opponent whenever he tries to maintain the status quo prevent each of the actors from being able to credibly commit to not building the weapon to extract that advantage. Since each also wants to avoid an outright defeat, both end up in a costly and useless arms race.

2 Stag Hunt: Trust and Mistrust

You might be tempted to think that perhaps the outcome of the Prisoner's Dilemma is due to the assumption that each actor wants to exploit the other when the other is cooperating. The next example shows you that this is not necessarily so. One possible objection to depicting the arms race as arising from a Prisoner's Dilemma configuration of preferences is that it seems to require the actors to be aggressive in the sense that they both prefer to compel the other to capitulate than to live with the status quo. What if this is not the case? What if they prefer to live with the status quo instead of trying to take advantage of each other?

To represent this situation, we only need to alter the ranking of the top two outcomes, as follows:

$$\text{Status Quo} \succ \text{Victory} \succ \text{Arms Race} \succ \text{Defeat} \qquad \text{(Stag Hunt Preferences)}$$

When both players share these preferences, the resulting game is called the *Stag Hunt*.[1] Since we are merely labeling the outcomes, the tabular representation remains unaffected (see Figure 1).

Unreciprocated cooperation is the worst possible outcome for each player, and mutual defection is the second worst outcome. However, both players prefer mutual cooperation to unilateral defection. Compare these preferences to the ones in the Prisoner's Dilemma: the only difference is that we have now flipped the top two preferences, meaning that no player has an incentive to defect when he thinks that the other is cooperating.

How would one play this game? The first thing to note is that which action A prefers depends on what he thinks B's action is going to be. If B is going to refrain from building the weapon, A can enjoy the status quo by not building the weapon or obtain victory by building it. Since he prefers the Status Quo to victory, his choice will be to cooperate, C. If, on the other hand, A thinks that B is going to build the weapon, then cooperation would result in defeat whereas defection would result in an arms race. In this case, A will choose to defect, D. Unlike the (strategically uninteresting) Prisoner's Dilemma, where each actor has a best strategy irrespective of what the opponent does, the Stag Hunt is more involved because each actor's best strategy depends on what he thinks the opponent is doing.

[1] The name comes from a problem posed by Jean-Jacques Rousseau, whose story (roughly) goes as follows. Two hunters must decide whether to cooperate, C, and hunt a stag together, or defect, D, and chase after a rabbit individually. If the both stalk the stag, they are certain to catch it, and they can feast on it. However, it requires both of them to stalk it, and if even one of them does not, the stag is certain to get away. If, on the other hand, a hunter goes chasing a rabbit, he is certain to catch one regardless of what the other one does. Assume that if the other one is also hunting for rabbits, the noise they both make scares the tastiest rabbits away and they can only catch stale hares with lower nutritional value. In other words, if you go after a rabbit, there is a slight preference that you do so on your own. Even the best rabbit is worse for a hunter than his share of the stag. There is only time to stalk the stag or hunt for rabbits, they cannot do both.

This means that in order to decide what A is going to do, he must predict what his opponent is going to do. The other actor, however, faces a situation analogous to his: *her* optimal action depends on what she thinks A is going to do. If she thinks that he will cooperate, then she prefers to cooperate as well. If she thinks that he will defect, then she prefers to defect as well.

Can we find a combination of actions for the two players that they would want to choose if their expectations about each other's behavior are correct? Consider the case where both are expecting to cooperate: $\langle C, C \rangle$. Since each player prefers to cooperate when he expects the other to cooperate, nobody would want to choose a different action, which means that their expectations of cooperation are correct.

Consider now a situation where A cooperates but the other player defects: $\langle C, D \rangle$. If A expects the other player to defect, he will not want to cooperate either. But then the other player has no reason to expect him to cooperate, which means that we should not expect players to settle on this combination of strategies. An analogous argument applies to the case where A defects but his opponents cooperates, $\langle D, C \rangle$.

Finally, consider the case where both defect: $\langle D, D \rangle$. Since each player prefers to defect when he expects the other to defect, nobody would want to choose a different action, which in turn means that the expectations of defection are correct.

We conclude that if both players wish to obtain the best possible outcomes for themselves, one of two things should happen: they will either both cooperate or neither will. With such two diametrically opposed outcomes, we really need to know which to expect.

Cooperation is best if you think the other is cooperating. These expectations are self-enforcing in the sense that *your* expectation of the other player choosing to cooperate rationalizes *your* choice to cooperate, which in turn validates *his* expectation that you will cooperate, which then rationalizes *his* choice to cooperate, and this in turn validates *your* expectation that he will cooperate, closing the circle of mutually supporting expectations.

Unfortunately, the exact same logic applies in the case of defection. If you think your partner will defect, you will defect as well, which validates his expectation that you will defect, which rationalizes his defection, which in turn validates your expectation that he will defect. Again, the circle is complete and we have a situation with mutually supporting expectations.

The question then seems to boil down to where we "begin" the circle of expectations. For instance, if we think one of the player expects the other to cooperate, we end up with the cooperative outcome. If, on the other hand, we think one of the players expects the other to defect, we end up with the non-cooperative outcome. So which expectation is more likely? Without knowing the actors and their relationship, it is impossible to say for sure.

One approach would be to say that both players know that the cooperative outcome is strictly better for both of them than any other outcome. It is definitely much better than the mutual defection outcome. This seems to imply that reasonable players should be able to see this, recognize the advantages of coordinating on this outcome, and do so without much difficulty. According to this line of reasoning, the Stag Hunt is not much of a social dilemma at all: the inevitable outcome would be mutual cooperation.

Not so fast! We could ask ourselves: if I were one of these actors, which is the *least risky* choice to make? That is, which choice gives me an outcome that leaves me least vulnerable to the behavior of the other actor? To answer this question, I need to figure out the relative

likelihood that I might be mistaken in my expectations about what the other actor is going to do.

Consider first the $\langle C, C \rangle$ equilibrium where both actors are expected to not build the weapon. Even though B prefers the Status Quo to Victory, we can assume that the difference between the two outcomes is not too great from her perspective; after all, the second-best outcome does involve a victory. If B's preference for the Status Quo over Victory is not very strong, A might worry that she is not going to put a lot of effort into playing C: B's **deviation loss** from choosing the non-equilibrium response D is just too small. But then A will really worry about getting his expectation about C correct: from his perspective, playing C but being wrong about B is disastrous because the outcome switches from the relatively palatable Status Quo to the worst possible, Defeat. When A cannot fully trust B to reciprocate cooperation, the prudential course of action might be to mitigate the risk of being saddled with the worst outcome by defecting and ensuring that at the very least one would get the Arms Race (or, if B does cooperate, Victory).

Thus, doubts about B's commitment to cooperation – doubts caused by the small deviation loss that B would suffer for failing to stick with cooperation – could increase the probability that A chooses to defect. Of course, B could go through the same reasoning and conclude that if A harbors doubts about her commitment, he might not cooperate. Note that this has nothing to do with B's actual intent to cooperate, which might be full. Instead, it is about A's beliefs about that intent, and B has no direct control over those beliefs. But now even a fully committed to cooperation B will start to worry: if A defects while she sticks to cooperation, she will be saddled with the worst possible outcome of Defeat. The prudential course of action would be to protect herself against such risk by defecting as well. *Thus, unsubstantiated doubts about B's commitment to cooperation have resulted in very real incentives for B to actually become less likely to cooperate!*

But, of course, A can go through the same reasoning process. He will know that his initial doubts about B give him an incentive to defect, which in turn increases B's incentive to defect, which now increases A's incentive even further. The process will continue like this, in a feedback loop, until both actors have virtually convinced themselves that the other will defect, and as a result both will. From this perspective, $\langle C, C \rangle$ is an unstable equilibrium: it can very easily by undermined by mistrust as very small initial doubts in even one of the players quickly cascade into mutual distrust that ends up preventing cooperation. The tragic irony here is that *both actors will be convinced that they are doing the prudent thing because they cannot trust the other.*

This logic leads us to expect $\langle C, C \rangle$ to not be a good rationalization for behavior despite being an equilibrium. Does $\langle D, D \rangle$ suffer from the same trust issues? If we assume that Defeat is much worse than an Arms Race, then both actors have very strong incentives to stick with their equilibrium strategies. Unlike the cooperate case, where (possibly inadvertent) deviation can cause little harm (because it would switch the outcome from Status Quo to Victory), deviation in the non-cooperative case can be disastrous (because it would switch the outcome from Arms Race to Defeat). This means that neither actor should worry that the other might deviate. Moreover, even if an actor does deviate, his opponent will actually be made even better off (because from his perspective the outcome would switch from Arms Race to Victory). Thus, there is no need to protect against mistaken expectations about the other's behavior. This reinforces the commitment to the equilibrium strategy and prevents

the negative feedback loop from even starting. From the risk perspective, then, $\langle D, D \rangle$ is a stable equilibrium. Distrust does not undermine it and trust cannot help break out of it. This makes mutual defection a much more convincing rationalization for behavior, and we can now understand why an arms race can occur even when mutual cooperation is, in principle, not merely possible but also an equilibrium. The problem is that the cooperative equilibrium is seriously undermined by distrust.[2]

This is a very pessimistic result: we both prefer the cooperative outcome to everything else, and this fact is common knowledge. And yet, even small amounts of doubt about the trustworthiness of the other player along with desire to protect oneself from being wrong about the other is almost certain to produce the second worst outcome for both us.

Another possible way to rationalize the baffling arms race is by noting that despite incentives to cooperate, the actors might worry about the consequences of being wrong in their expectations about the behavior of their opponent. When they have small opportunity costs of sticking with the cooperative behavior, deviations from that behavior become "too easy", and worries increase. Lack of trust can cause actors to choose the prudent course of action that minimizes their vulnerability to having been wrong about trusting the other, and as a result they can end up in a costly and useless arms race.

3 So, What About that Arms Race?

The logic of the arms race in a SH-like scenario is fundamentally one of mistrust, risk-aversion, and prudential reasoning. The logic of the arms race in a PD-like scenario is one of desire to exploit the other side's cooperative effort combined with a desire to avoid being saddled with the worst possible outcome. In this sense, the Stag Hunt probably captures the dynamics of fear-induced hostility much better than a Prisoner's Dilemma.

The advantage of a SH-like situation over a PD-like situation is that the social dilemma is solvable in principle in the first case but not in the latter. For instance, if we manage to coordinate expectations and attain a level of trust between ourselves, we will cooperate in SH but still will not cooperate in PD. The cooperative outcome can be sustained in equilibrium in SH but not in PD, which implies that one possible solution to cooperation failure in SH is to work on expectations.

In international politics, one cannot know the intent and motivations of one's opponent (or partner). We cannot peek into the heads of decision-makers to verify that they do not intend to attack us, which is (of course) what they usually claim. Intentions are not only unverifiable, they are volatile. Changing governments, the particular mood of the leader, or many other factors may change the evaluation of the desirability of attack on a moment's notice. This is why states normally do not rely on intentions, they are forced to *infer* intent from *observable* capabilities and behavior.

This is where suspicion comes into play. If I cannot be certain that my opponent has no intention to attack me, I must admit the possibility (however small) that he might do so. Since being defeated is the worst possible scenario for me, prudential reasoning might lead

[2] In case you are wondering, $\langle C, C \rangle$ is called a **risk-dominated** Nash equilibrium. Work on evolutionary models has shown that natural selection might lead to strategies that diverge from risk-dominated equilibria toward the risk-dominant one; in this case, $\langle D, D \rangle$. Work by Harsanyi and Selten has shown that most games have a unique risk-dominant equilibrium.

me risk losing the cooperative outcome in favor of securing, at the very least, a costly preservation of the status quo. So I build some weapons to guarantee my security. Unfortunately, my act of increasing my security immediately decreases the security of my opponent. He would reason as follows: "I was almost sure that he did not have hostile intent but now I see him arming. I know he claims it is purely for defense but is that so? Perhaps he intends to catch me unprepared and defeat me? And even if that is not so, he clearly does not trust me enough or else he would not have started arming. I would like to reassure him that I can be trusted but the only way to do so is to remain unarmed, which unfortunately is very risky if he does happen to have aggressive intent. So I better arm just to make sure I will not have to surrender in that eventuality."

My opponent then arms as well, which makes me even less secure. We both have matched each other in armaments, the status quo survives, but we also learned that we cannot trust each other not to arm. Because we cannot observe intent, we can only see the arming decision which could be because the other side is afraid or it could be because the other side is aggressive. In other words, neither actor can be sure about the preference ordering of the opponent: is it SH-like or PD-like? Moreover, when an actor with SH-like preferences – and thus no intent to exploit the cooperation of the other – sees the other arming and possibly claiming it's out of fear, it might be very difficult to believe that actor *precisely because one does not harbor any aggressive intent*. It is hard to put oneself in the other's shoes, and when one is innocent of sneaky designs, one is more prone to conclude that other one cannot possibly believe their own hype, and as a result infer that the other side does, in fact, have such designs that they are trying to cloak as self-defense. Without drastic reassurance by the opponent this suspicion, then, can lead one down the path of self-preservation too, despite being initially willing to reciprocate cooperation. Reassurance being too risky, we opt for the prudential choice and continue arming, further increasing the suspicion and hostility. The process feeds on itself and rationalizes the non-cooperative outcome, just as in the original Stag Hunt story. The process, in which small doubts lead to defensive measures which increase the insecurity of the opponent, who reacts with defensive measures of his own, which increases my insecurity and as well as my doubts leading to further defensive measures on my part, is called the **Security Dilemma**, and it is very similar to the Stag Hunt scenario.

Notice that once the suspicion starts, it is in the interest of the players to restore trust and get the cooperative equilibrium. Unfortunately, trust can only be restored if one of the players decides to take the risk and plunge into unilateral disarmament. If his opponent turns out to have a SH preference structure (prefers the status quo without arms to victory), then this gesture would be reciprocated and the players could potentially go to a stable cooperative solution. If, on the other hand, one's opponent turns out to have a PD preference structure, then one risks defeat. If one suspects that the opponent has PD preferences or if one's opponent is so suspicious that he would ignore the gesture, no player would make the necessary first step to achieving cooperation.

What model you think represents the Arms Race problem best depends on what you think the structure of the preferences is. If you think of the Arms Race as a Prisoner's Dilemma, you would not recommend trust-building and risky unilateral actions: the opponent is sure to ignore anything you say and would not reciprocate restraint because exploiting your weakness is preferable to cooperation. If you think of the Arms Race as a Stag Hunt, on

the other hand, you would recommend trust-building, and you might even recommend a dramatic unilateral gesture that runs serious risks but that can persuade the opponent of your peaceful intent. (We shall see how precisely this type of gesture by the Soviet Union was the catalyst for ending the Cold War.)

These illustrations underscore the major reason for doing this abstract analysis. Once we learn to recognize the equivalence of different strategic situations, we can apply the insights from a model describing one of them directly to another without even having to build a model to represent it. In this course, our goal is to study a series of games to build our intuition about what types of situations seem to occur that concern national security. Once we begin recognizing the similarities (strategic equivalence) between different situations, we can apply our insights to analyze them without actually having to construct explicit models. We shall see that the abstract games tell us quite a bit how to deal with adversaries as disparate as the Soviets, Saddam, or terrorists!