U.S. Foreign Policy: Theories & Strategic Choice

Professor Branislav L. Slantchev

July 2, 2014

1 Theory and Explanation

In everyday use, the word **theory** has very unfortunate connotations. It is often taken as equivalent to *speculation* or *idea*, as in "human-caused climate change and evolution are just theories" (with the implication that they lack the empirical evidence necessary to support their claims), or to **hypothesis**, as in "I have a theory where my missing car keys might be" (with the implication that it is a conjecture that is subject to empirical verification). In science, however, the word **theory** refers to a well-articulated *explanation* of some phenomena that is also *well-supported* by empirical evidence. In this sense, it is well beyond mere speculation: it tells us *how* the phenomenon in question works, and this explanation generates a variety of testable propositions, or *hypotheses*, that can be, and have been, evaluated empirically through experiments and observation. The overwhelming evidence supporting these hypotheses is then taken to indicate support for the theory itself. Although no theory is ever "proven" in the sense of a mathematical theorem, at some point the evidence in support is so overwhelming that only an insane person who is made aware of it would deny it. Heliocentrism is "just" a theory, but I dare you to find any reasonably educated person who would deny that the Earth orbits around the Sun.[1]

The opposite of theory is not fact but mystery.

1.1 Causal Mechanisms

Our use of the word *theory* is not going to be anywhere near the scientific ideal but it will be much more demanding than the everyday use. For us, a theory must provide

[1] It is astounding that in 2012 one in four Americans still believed that the Sun goes around the Earth. See Table 7-8 in the report by the National Science Foundation, `http://www.nsf.gov/statistics/seind14/content/chapter-7/c07.pdf`. Europeans fared even worse: in 2005 one in three failed this fundamental astronomy question. I have no data on the distribution of incorrect answers by educational level but it is stunning that one can get out of high school and still believe the equivalent of the Earth being flat and resting on the back of a world turtle, with larger and larger turtles all the way down.

an explanation: a **causal mechanism** that tells us how some variables interact with each other to produce the outcomes we seek to understand. Notice that the selection of theory depends on the target: what is the question we seek to answer? The question is usually something that confounds our expectations, something that we do not understand, and so something that needs to be explained. Theory provides the answer in the form of a mechanism that establishes a causal chain between the variables and the outcome.[2]

Let us start with some historical phenomenon that we might wish to understand. The most obvious problem with history is that there are too many variables one could potentially look at. Which are important and which can safely be discarded? How do we decide? The answer is that we need a "guide" to selecting variables. This is what theory does: it tells us how some variables interact with each other to produce the phenomenon in question (which is another way of justifying the need to look at these, and not other, variables).

Consider a hypothetical example. Suppose we observe a statistical correlation between war initiation and high unemployment. Our hypothesis would be that high unemployment causes wars of aggression. We now need a theory that provides the mechanism that links the explanatory variable (unemployment) to the explanandum (war of aggression). We can hypothesize that high unemployment (a) causes social unrest that could be channeled toward an enemy, (b) causes governments to expand employment in armament industry — reduces unemployment and is justified by attributing hostile intent to enemy, (c) causes governments to search new markets to encourage producers to hire workers — aggressive foreign policy, (d) gives rise to populist leaders who are more aggressive in foreign policy. We could now use this theory to check whether the cause has the hypothesized effects which in turn produce aggressive wars. But we could also continue to refine the theory by opening up (d): why would high unemployment bring populist leaders to power? We could theorize that (d-1) the natural clientele of populist is more likely to vote (or engage in political behavior) when its opportunity costs are low — which they will be when unemployed since there is no income to forego; (d-2) populists are more likely to promise instant solutions to unemployment; (d-3) populists offer to punish those that the unemployed believe to be responsible for their plight. Again, each of these hypothesized effects can be checked against data. But we do not have to stop there: we could want to know how those "guilty" for the plight of the unemployed are identified and punished. We might hypothesize that (d-3-1) the wealthy would be worried about the security of property rights and so would be willing to strike deals with the government in which they relinquish some of their wealth in return for protection — redistribution toward the unemployed; (d-3-2) they might support

[2]In this way our theories are not merely instances of abstract generalized thinking, as in "music theory" or "art theory" or "literary theory", but instead specify models/mechanisms that ultimately yield testable propositions (hypotheses) that can be subjected to experimental and quasi-experimental evaluation.

the leader in aggressive foreign policies that blame the enemy in an effort to deflect attention from themselves. These hypothesized effects would predict that high unemployment would be associated with some internal redistribution of wealth and with propaganda vilifying an external enemy. The latter can lead to crisis escalation and, possibly, war.

1.2 Rationalist Explanations

For a mechanism to be of any use, it has to go beyond providing a list of variables and effects. Since the phenomenon we are interested in here (war) is ultimately produced by the behavior of people, a mechanism should be anchored in individual behavior. In other words, it should tell us why the relevant agents acted in particular ways in given contexts. But how do we understand individual behavior — generally, we do so by **rationalizing** it. That is, we take the observed behavior we seek to understand, and then attribute some preferences and beliefs to the individual that engaged in it such that this observed behavior is expected to contribute to the welfare of that individual as defined by his beliefs and preferences. We assume that individuals are "rational" in the sense that their actions are purpose-driven so that individuals tend to behave in ways that are supposed to enhance their well-being. How individuals define well-being and how they analyze their environment depends on their preferences and beliefs. The actions they can choose from depend on the context in which they act and the information they have; that is, on institutional and informational constraints. An idealized "rational agent" always chooses the optimal course of action, with "optimal" defined as the course most likely to deliver on the desired goals.

All of this is purely hypothetical: we use observed behavior to infer preferences and beliefs that make this behavior optimal given the constraints. We then explain the behavior by saying that it must have been the result of the purposeful pursuit of the goals we attributed to the individual. This sounds suspiciously ungrounded in reality, and it would be without some means of testing the various connections this mechanism requires in order to make the causal chain work. The virtue of having the theory is that it tells one which variables to look at, how they should change, and what their effects should be — all of this can be subjected to empirical testing (observational or experimental). We could attempt to ascertain the preferences and beliefs the relevant individuals had to see how closely they match our assumptions about them. We can go further and ask whether it is *reasonable* for the individual to have held these beliefs given the information this individual had at the time. We would also attempt to analyze how closely the constraints we assumed are matched by the context in which the individual had to act. Matching closely these factors would give us confidence that the mechanism we postulated is, in fact, explaining behavior. We could say that we understand it because we can rationalize the behavior of the relevant individual with some confidence.

Why focus on rationalist explanations? For starters, people *want* to be rational in the sense we've been using the word. They want to have "good" reasons for their behavior, which is why they often "rationalize" them after the fact by pretending to have had goals or beliefs that would make their behavior reasonable. More importantly, we rely on this sort of reasoning all the time when we want to make sense of the behavior of others and when we want to predict how others will react. In fact, when we fail in these predictions we are apt to characterize the surprising behavior as irrational.

This is not to say that "irrational" behavior must be unintelligible. For example, strong emotions might short-circuit decision-making and cause individuals to rush into actions they otherwise would not have. Shame might cause one to commit suicide; fear might cause another to jump out of a burning building. Desire for revenge might motivate actions that are exceedingly costly personally with little objective benefit even if they succeed. (In these, however, some element of ratiocination might remain if the individual still chooses the course of action that is most likely to cause the desired result.) Weakness of will is often behind failure to lose weight or, in some cases, quit smoking. Wishful desires bias belief formation, causing individuals to stop searching for better solutions or more information, or to discard information contrary to their desires. There are many other psychologically motivated biases in decision-making that might produce actions that fall short of the optimal. Going into psychiatric explanations, there are also the various obsessions, phobias, delusions, and so on. Any of these can make behavior intelligible, so why should we privilege rationalist explanations?

The main reason for that is that irrationality can "explain" too much too easily. People often attribute puzzling behavior to irrationality when in fact it could be perfectly rationalizable by factors they fail to consider. Take, for example, the Marxist hypothesis about *false consciousness*. According to Marxism, the proletariat does not have a shared interest with the capitalists in policies that enhance the well-being of the latter (because this could only increase the exploitation of the former). An example of such a policy, at least according to Lenin's view, would be "imperialist wars," that is, wars fought by capitalist societies over access to markets and colonies for raw materials. Since it is precisely the members of the proletariat who die is soldiers in these wars but only the capitalists stand to reap the profits, it is in the workers' interest not to support such wars. When the First World War broke out, many Marxists in fact expected the masses to recoil from service. Unfortunately (for theory and for the masses), the opposite happened — not only did proles from one country enlisted in their armies, in many cases voluntarily, but they did not seem particularly reluctant to kill "fellow" proles from other countries with whom they supposedly shared interests in overthrowing capitalists. This was a clear divergence from behavior that class interest would dictate. The theory was "saved" by the notion of "false consciousness" according to which the ideological control of society by the bourgeoisie and nobility has blinded the proletariat to its true class

interests. The proles either do not know that interest (because, for example, religion tells them what the "natural order of things is") or they do but choose to disregard it because they are promised to enter the ranks of the privileged. Whatever the reason, the proletariat's acting against the interests postulated by the theory is "explained" by amending the theory to essentially argue that the proletariat is deluded. (A much simpler explanation would have been that the theory is wrong.) Thus, according to Marxist theory, the proletariat will act in its own interest except when it does not. Observationally, when we observe workers unionizing and striking, the theory is supported because it is in the interests of workers to force the capitalists to share in the surplus their labor creates. When we observe workers acting in concert with capitalists to thrash other workers and their capitalists, the theory is supported because they are acting out of false consciousness.

There is no possible behavior that the workers can engage in that can *falsify* the theory, even in principle. This means that we have to take the theory on faith — there simply exists no sort of evidence that could potentially disprove it. But if the theory were wrong, how would we then know this? In the above example, we could not. This renders the theory useless as an explanatory device: everything that does not conform to one postulate conforms to another in the same theory. We shall require our theories to have a property known as **falsifiability** — meaning that if the theory is false, then there does exist some sort of evidence we can obtain either by observation or by experiment that would demonstrate that. Without false consciousness, Marxism is falsifiable — the evidence of workers failing to act in their class interests would show that the theory is wrong. With false consciousness, Marxism is unfalsifiable since all evidence is consistent with the theory. It is not that one should discard a theory at the first sign of non-conforming evidence — that would be naïve. One can always seek to amend the theory to account for any new evidence in addition to all the evidence it could previously handle. However, when such an amendment goes too far — like false consciousness does — it can render the resulting theory unusable.

Rationalist explanations are in a way *minimalist explanations* because they are the ones most readily falsifiable. This makes them particularly suitable for hypothesis testing, which allows for accumulation of knowledge and verification. Explanations that rely on irrationality do not have to be non-falsifiable (although some of them are). The problem is that they are too convenient and so might lead to ignoring the actual mechanism. It is all too easy to say "oh well, he acted out in anger" instead of searching for other causes explaining puzzling disregard for one's own safety. In fact, the ability to mimic irrational behavior for rational reasons should give one further pause before reaching for such explanations. If an individual "acts crazy" for the purpose of convincing others that he is crazy (meaning that they cannot rely on usual cost-benefit reasoning to predict how he would act), he is not really crazy — provided the others believe him and adjust their behavior accordingly. He is cunning, he is strategic, he is supremely rational in choice of action given his

goal.

To give a specific example, how are we to understand the 2003 Iraq War or, more specifically, how are we to understand the behavior of Saddam Hussein? In the light of the outcome of the 1991 war over Kuwait, the subsequent degradation of the Iraqi armed forces, and the continued improvement of the US military, it would appear nearly certain that a war with the US would have inevitably ended in the overthrow of the Iraqi dictator. So why pursue policies that clearly tilted the US toward war and, more importantly, why persist after it became clear that the US will, in fact, invade? One answer is that Hussein was irrational, so these calculations simply did not enter his mind. He might have put his faith in God or in his own genius. This, however, sounds more like a label than an explanation. One could instead argue that Hussein made a mistake because he was misled as to the true state of his military by advisors who were too afraid of him to reveal just how much it had deteriorated. This would have given him false optimism and encouraged him to resist. (Similarly, he might have expected the US to be incapable of forming a grand Coalition of the 1991 type — which was correct — and thus be reluctant to fight on its own — which was incorrect.) This explanation would rationalize his behavior by showing that it was reasonable given the information he had at the time. An even stronger version would argue that even while there was no uncertainty about the military outcome of an American invasion, there was far more uncertainty as to the fate of subsequent pacification — would the Americans have the stomach to stay and fight for years on end an enemy that mingles with civilians and that cannot be readily identified and defeated in pitched battle? If Hussein could survive the initial onslaught and then organize national resistance to the occupying forces, then resisting the US makes sense especially if failure to do so would expose the weakness of the dictatorship and make Hussein's overthrow nearly certain. This type of explanation rationalizes his behavior by showing that he took a calculated risk, a risk that actually made sense despite the overwhelming military superiority of the United States. Even though he eventually failed, the behavior had been reasonable. Which of these (or the myriad alternative) explanations is valid depends on the assessment of the facts and how closely they track the connections identified by the various theoretical mechanisms.

1.3 The Map Analogy

A final word about theory: it is *not* a full description of reality. It cannot be: the closer it gets to reality the less useful it becomes as a means of understanding that reality. The power of theory is in that it abstracts away from the complex real world and attempts to reduce its vastly complicated interrelationships to a small set of manageable variables and connections. In this, a theory is like a map. How useful this simplification is depends on the purpose (which determines how much detail you can omit without producing a useless map) and how good the theory is

(it includes all the variables it has to in order to produce reliable predictions about their effects). Neither of these is really known *a priori*, so each theory is essentially a bet that its particular formulation would be useful.

Each theory is then "valid" while it continues to be useful. It is not discarded when one encounters contradictory evidence, especially if there exists no alternative that can take its place. The theory can be modified to account for that new evidence although care should be taken that the adjustment is not ad hoc, meaning that the new version should handle what the old theory could plus the new evidence plus whatever new hypotheses it gives rise to. It is a tough order for a new theory to pass, which is why we have long used theories known to have "holes" in them — Newtonian physics is one example, Ptolemaic astronomy is another — they are good enough for most purposes and there was no viable alternative — until, that is, Einstein's theory of relativity and Copernicus' theory of Heliocentrism.

Going back to our map analogy: how useful would it be to have a map that is an exact representation of reality? For starters, it would be impossible to create one: it would have to be as large as the world it represents. OK, so the first "compromise" would be to reduce it to manageable proportions, say 1 to 5,000 (1 cm to 50 m), which would be useful for a walking map. Obviously, going that small means discarding a lot of detail. So what can we let go? It depends on the purpose of the map. If we want a walking map, then we should retain roads, paths, trails, some information about the terrain, and relevant markers. If we want a driving map, we need roads but can omit foot trails, we might want to include gas stations and rest stops, and so on. A walking map would not be useful in a city if we wish to use the bus, and a map of the bus routes would not be useful if we need to use the subway. In fact, anyone who's ever looked at a map of bus routes or subway lines would be familiar with the highly idealized schematic representation of reality they represent — nice straight lines with nice junctions at right angles and often stations equidistant from each other — in short, very little of reality has made it onto these maps. Yet they are far more useful for those trying to utilize the respective modes of transportation than a highly detailed physical map of the place or a nicely illustrated map of tourist attractions.

Theories work the same way: purpose determines scale and simplification. The trouble is that unlike a map — where purpose fairly clearly dictates content — no such useful guide exists for theories. We have to formulate them, produce tentative hypotheses, proceed to experimental and observational verification, then reformulate as necessary. No theory is ever final (and that's a good thing) — theories are always the best we can do with the knowledge we currently have. This makes them tentative and subject to revisions. Theories that have withstood the test of time acquire the special status of scientific "truth" because we have yet to uncover disconfirming evidence. But this "truth" is not absolute, it is not dogma. It is no more nor less than a reflection of what's possible in our state of the world.

2 Strategic Choice

In order to organize our thinking about foreign policy, we must decide what it is that we want to study and what assumptions we want to make to simplify reality sufficiently to make it comprehensible. As I noted above, our explanations of foreign policy must necessarily boil down to arguments about why particular actors took certain actions, and these arguments must, on average, hinge upon the assumption that somehow each actor was trying to achieve some desired goals. Since no actor, not even the President of the United States, is powerful enough to simply impose its preferred outcomes on others, the defining characteristic of international relations (and so, foreign policy) is the **interaction** among various actors, and it is this interaction that we shall study. At the most abstract level, we must distinguish three components: (i) the actors, (ii) the environment in which they act, and (iii) how outcomes are produced from the actions.

2.1 The Actors: Preferences and Beliefs

Here are some examples of different actors in whose interaction we might be interested: states fighting a major war, United Nations engaged in peacekeeping operations, governments of two states negotiating a trade treaty, the ministries of a country seeking accession into the European Union, State Department and Department of Defense struggling for control over foreign policy, General Motors and Ford lobbying the government for protection against "unfair" foreign-trade practices, French farmers dumping grapes to protest agricultural policies of the EU, individuals engaging in terrorism.

It should be evident that we are not interested in fixing some particular level of social aggregation as the unit of analysis. That is, we do not want to say that we shall investigate relations between states only, or between leaders of states, or even between organizations within states. International relations are far less conveniently structured than this, and we shall have to account of various different types of actors getting involved.

To deal with this complexity, we shall use an abstract definition of an actor. An actor has two attributes: **preferences** and **beliefs**.

To say that an actor has preferences simply means that it can rank order different outcomes according to some criterion or criteria. For example, consider the situation with Iraq and suppose there are six possible outcomes: (i) Iraq provides acceptable proof of dismantling of its WMD programs, (ii) Iraq agrees to dismantling whatever is left of these programs under international supervision, (iii) Saddam steps down as Iraq's leader, (iv) the United States invades Iraq and wins, (v) the United States invades Iraq and loses, or (vi) the US does nothing.

The United States is an actor that has a specific preference ordering. That is, it ranks these alternative outcomes in some rational way. Similarly, we can designate

the State Department, or Saddam, or President Bush for that matter as actors, and they all will have their own preference orderings.

The other attribute of an actor is the beliefs it has about the preferences of other actors. Again, since we are interested in interaction among actors, we want to know how these actors will behave, which in turn depends on what they think others will do. To form an expectation about the behavior of other actors, it is necessary to have some belief about what preferences the other actors have. For example, we might be uncertain about whether Saddam's preferences are such that he prefers (i) to (ii) above, but we can hold a belief about the likelihood that it is the case. When actors are uncertain, as it is usually the case because they seldom possess complete information, beliefs are crucial to the choice of action.

Thus, we shall study the interaction among actors, where actors are defined by two attributes, their preferences and their beliefs.

2.1.1 Unitary and Composite Actors

It is important to understand that actors that we can profitably treat as single "individuals" at a high level of abstraction can themselves be composed of other actors at a somewhat lower level of abstraction. For example, in some contexts, it might be appropriate to define the United States as the actor and postulate some preferences over the risky alternatives. This could be a useful shortcut, and historians often employ it, in some situations: for instance, it might not be too distorting of reality to treat the United States as an actor whose preferences opposed the expansion of Soviet communism during the Cold War. In other contexts, however, this shortcut might be distorting: for instance, we might wish to analyze how the United States would respond to some particular aggressive move by the Soviet Union. Reasonable people can agree on the goal — preventing the success of this move — but disagree about the appropriate course of action. This disagreement can arise because of different political priorities, beliefs about "how the world works", or even organizational and bureaucratic issues.

Here we would need to "disaggregate" the United States into a composite of several relevant actors. But how do we know what these actors have to be? To answer this question, we need to know a bit more how the U.S. foreign policy decision-making process is organized. We shall study this in some detail very soon, so for now let us assume that the important individuals would be the President, the National Security Advisor, the Secretary of State, the Secretary of Defense, and the Chairman of the Joint Chiefs of Staff. We can take these individuals as representing the preferences of the respective organizations they head, which means that they might have very different ideas how the Soviet action might need to be handled. For instance, the JCS chairman might prefer to respond militarily with an action that has the highest chances of success; e.g., a ground invasion. The Defense Secretary might agree with the assessment of the likelihood of success but

might be more sensitive to the costs the various actions entail. He might prefer to opt for a much less expensive strategy — e.g., massive air strikes — that might have a smaller probability of success. The State Secretary might be worried about the fallout of using a military option without attempting a diplomatic solution first. He might prefer to delay the military response until allies could be consulted and the opponent given an opportunity to retreat without an overt confrontation. The National Security Advisor might believe that the Soviet move, while seemingly aggressive, is not actually all that threatening and that even if successful it would not really damage American interests. He might be opposed to any military response but also to any diplomatic intervention which might disturb the allies unnecessarily and give the opponent an opportunity to score points by defying the negotiation attempts. He might think that the appropriate course of action is to do nothing at all and simply ignore the Soviet move. The President might agree with absence of a real threat, but worry about the impact of inaction on the American public; he might believe that the public would never forgive him for failing to resist aggression. Thus, he might want to do something, and that something would have to be more than "merely talking" about a diplomatic solution but definitely less than immediate military action; he might, for instance, decide on a forceful non-military response like a naval blockade.[3]

In order to explain the foreign policy choice of the United States government in this scenario, the theory cannot treat the United States as a **unitary actor**. Instead, it will have to incorporate knowledge about the decision-making process at the highest level of government to model the United States as a **composite actor** whose preferences and beliefs are somehow determined by the preferences and beliefs of the five unitary actors we identified. At this point, the theory will confront two issues. The first is merely a repeat of the unitary actor problem we just encountered: even though the Secretary of Defense is an individual, it does not follow that he has to be modeled as a unitary actor; after all, he is the head of a vast, and fairly complex, bureaucratic organization that, at a minimum, comprises structures designed to deal with the three main branches of the military: the Army, the Navy, and the Air Force. When the President asks for advice, the Secretary would focus on the level of forces and manpower necessary to pursue various goals. As the head of this agency, he might concerned about the appropriate balance among the various branches, their degrees of readiness, and cost effectiveness. He might wish to pursue organizational goals that involve promoting some particular technology at expense of others: e.g., a new stealth bomber instead of more tanks. This organizational goal might bias him in favor of air strikes (that would demonstrate the capabilities of the bomber, and so earn him even more support from the Air Force) and against ground invasion (that would expose the shortage of tanks he had

[3] Students familiar with the 1962 Cuban Missile Crisis will, of course, recognize that this hypothetical example is not fabricated out of thin air.

created, and so open him to criticism from the Army). To fully understand the preferences of this actor, one might have to disaggregate it in turn, taking into account the bureaucratic organization of the Defense Department and the individual preferences of the various Secretaries and Under-Secretaries in it. Analogous issue might prompt us to disaggregate the Secretary of State, and so on. Clearly, this type of analysis can become extremely involved and so detailed that it would be nearly impossible to follow. For practical purposes, disaggregation stops at the highest level of abstraction that allows us to make meaningful predictions about the behavior of the composite actor. As before, purpose determines scale and simplification.

2.1.2 Preference Aggregation in Composite Actors

The second issue is the problem of **preference aggregation**. Even though we identified the five actors — which for now we shall treat as unitary — that are of special relevance for the formulation of foreign policy, we have not specified how their preferences and beliefs are aggregated into preferences and beliefs of the composite actor the United States. It could be, of course, that the President acts like a dictator and just implements the action according to his own preferences and beliefs. As we shall see, however, even if the President is ultimately responsible for the final decision, that decision will invariably be shaped by the opinions of those around him. This influence can be informal: the other actors seek to obtain agreement with their preferred action through a process of deliberation and persuasion. The influence can also be formal: the President takes the action that garners the majority vote. Different Presidents will employ different styles of decision-making, and it can run the gamut from near dictators who ignore advice to first-among-equals who carry out the wishes of the majority. They will also surround themselves with different types of individuals, some preferring the company of those whose preferences are not too dissimilar from theirs, and others valuing diversity of opinion.

Suppose that, after intense deliberations all five agree that neither land invasion nor doing nothing are desirable options. They still disagree, however, about the relative merits of air strikes, blockade, and diplomacy. Let's suppose, for the sake of example, that their individual rankings are as follows:

President	Advisor	State	JCS	Defense
blockade	blockade	diplomacy	air strikes	diplomacy
air strikes	air strikes	blockade	diplomacy	air strikes
diplomacy	diplomacy	air strikes	blockade	blockade

Table 1: Preference orderings of five unitary actors for the composite United States.

Since they cannot persuade each other beyond this, the President decides to use pairwise majority voting. He first asks everyone to choose between blockade and

diplomacy. Since three of the five actors prefer diplomacy to blockade, diplomacy is the winner. The President then asks everyone to choose between diplomacy and air strikes. Since three actors prefer the air strikes, the air strikes is the ultimate winner. It appears that the United States prefers air strikes most, followed by diplomacy, followed by blockade. The Chief of the JCS will be happy, but the Secretary of State is distinctly unhappy with this.

Suppose the State Secretary managed to persuade the President to redo the voting but start with the choice between air strikes and blockade. Since three actors prefer blockade to air strikes, the majority winner is blockade, which is then paired with diplomacy. But since three actors prefer diplomacy to blockade, the ultimate winner is diplomacy, which the State Secretary likes a lot. It now appears that the United States prefers diplomacy most, followed by blockade, followed by air strikes. It should already be troubling to you that a "mere technicality" of switching the order of voting has altered the preferences of the composite actor.

It gets worse. The President, who is now saddled with his least preferred option, has warmed up to the idea of agenda manipulation and decides to redo the voting. He asks everyone to vote on air strikes and diplomacy first. Since three actors prefer air strikes to diplomacy, the winner is air strikes, which is then paired with blockade. Since three actors prefer blockade to air strikes, the ultimate winner is blockade, just what the President wanted. It now appears that the United States prefers blockade most, followed by air strikes, followed by diplomacy.

Thus, depending on the order in which alternatives are considered, using majority voting to determine the preferences of the composite actor from the logically consistent individual preferences of the constituent unitary actors gives us logically inconsistent results, known as **preference cycles**. The United States prefers blockade to air strikes, air strikes to diplomacy, and diplomacy to blockade. These preferences are logically inconsistent because logic dictates that if one prefers blockade to air strikes and air strikes to diplomacy, then one should prefer blockade to diplomacy as well (preferences should be transitive).

The problem with preference cycles is that they make theories unfalsifiable because *every* choice is consistent with the preferences of the composite actor. But if every choice is "rationalized" by these preferences, then we cannot understand why any particular choice was made. It seems that any theory that seeks to rationalize behavior based on preferences is doomed from the start.

2.1.3 The Need to Consider Institutions

Or maybe not. In fact, our simple example above already suggests one way in which the preferences of the composite actor can be guaranteed to be consistent. If the President acts as the **agenda-setter** and decides the order in which options are brought up for a vote, then he can ensure that the preferences of the United States are exactly the same as his own even though they were ostensibly created

by majority voting. Thus, the agenda-setter can not only avoid cycles in aggregate preferences, but can usually ensure that the voting outcomes are very close to his own preferences. This gives agenda-setters considerable power, of course, which is why these formal positions are so desirable when the institutions allow for them. In our case, the President's elevated rank might informally designate him as the agenda-setter even when there is no formal voting rule in the group of decision-makers he is consulting with. This ability might, in fact, allow us to treat the United States as a unitary actor after all, except in this case its preferences would be those of the President. If, on the other hand, we were interested in the decision-making of another type of government, say a military junta composed of several generals who make collective decisions using majority voting, then we might be able to restrict attention to the general with agenda-setting powers.

The American government system of checks and balances, however, ensures that when it comes to foreign policy, the President might find himself at loggerheads with Congress. The ultimate action the government takes will be based on preferences created by aggregating the preferences of the executive and legislative branches. Congress itself is a very complex institution whose members have to deal with a great variety of possibilities, making the possibility of preference cycles quite distinct. Congress, however, has many rules and practices that eliminate that possibility altogether. Among these institutional features are: (i) the rules of order, which might limit the opportunities for defeated proposals to come back; (ii) reversion points (e.g., preset spending allocations in a budget), which automatically select an alternative if no proposal receives enough votes to decisively defeat it;[4] (iii) adoptions of winning alternatives as reversion points, which makes it exceedingly unlikely that voting would cycle back to the original; (iv) committee systems that limit the number of alternatives considered, amendment rules that require that any changes be germane to the committee proposal, or rules that limit the amendments themselves to those proposed by the committee;[5] (v) vote-trading practices, which allow a member to exchange a vote on some issue of interest to others for the others' votes on an issue of interest to the member (this allows for the formation of stable winning coalitions); (vi) parties, which restrict the domain of admissible preferences by enforcing party discipline on the members.

The institutional constraints and practices might appear arbitrary and might have somewhat undesirable consequences (e.g., logrolling can produce vastly inflated budgets, and party discipline might polarize Congress resulting in policy deadlock), but they are necessary evils because they impose structure that can induce stability in context where decisions are made by majority rule. This is why shall often

[4]This also works in law, where the current law stands unless the court explicitly overturns it; the principle of *stare decisis*.

[5]Under the Closed Rule in the U.S. House of Representatives, no amendments may be offered other than those recommended by the committee itself, which further restricts the range of admissible preferences.

have to consider the institutions in which policy-makers operate, not merely their (imputed) preferences. This is also why we will need to study the process of foreign policy formation in the United States more closely.

2.1.4 The National Interest

The problem of preference aggregation is much more pressing than our abstract examples might suggest. Consider, for example, the ubiquitous notion of **national interest**, in whose name political leaders and groups purport to act. There are two things here that we should be careful about:

1. How is the national interest determined, and
2. How is the most appropriate action chosen given that national interest?

That is, as a society we probably need to agree on what our common interests are, and once we agree on that, what the best ways to achieve these interests would be.

You have all read history books and are aware of stuff you see on TV. Not a single day goes by without some pundit pontificating on air or in print about the current crises in Iraq and with North Korea, not to mention the perennial Arab-Israeli conflict in the Middle East, the economic difficulties of Latin America, the AIDS epidemic devastating Africa, or the corruption scandals rocking Europe.

All of these discussions are invariably framed in terms of preferences of the participating actors. Historians, journalists, economists, and political scientists are all intensely interested in these preferences because we all look for explanations of behavior by assuming some consistent pursuit of self-interest by these actors. Whether in trying to divine Saddam Hussein's preferences or those of the United States, we all resort to an appeal of instrumentally rational behavior to explain what goes on. ("Instrumentally rational" refers to the assumption that people pursue actions consistent with their goals. That is, people will not willingly hurt their own interests.)

For simplicity, many analysts take the state as the unit of analysis when it comes to important international events. So we talk about a Second Persian Gulf War between America and Iraq, or a crisis between the U.S. and North Korea, or bargaining for more money between Turkey and the U.S. In other words, we often take the state to be the important actor whose behavior we want to explain. It is in this context that you frequently hear the much abused and maligned term "the national interest." But what is it?

There are several possible ways we can approach the problem, and all of them have been used in international relations theory:

- Objective interest, which overrides all other concerns whether states realize that or not. For example, realism postulates that state survival is the most important national interest and all other goals are subordinated to this one.

Liberals tend to argue that the world is not such a dreadful place and that economic well-being is the most important national interest.

- Expression of elite choice. In this view, elites have specific interests that they pursue through the state apparatus, to which they have better access than ordinary people. Elites then "sell" these policies to the rest of us, inducing our choices to conform to their preferences. This works both for democracies and non-democracies (authoritarian or totalitarian regimes).

- Expression of people's choice. Proponents of democracy argue that the national interest is simply an aggregation of individual preferences. That is, each and every one of us has his or her own preferences. In a democracy, we would then use some aggregation mechanism, usually voting, to arrive at the social preference.

Of course, there is no such entity as a state when it comes to preferences. States do not have preferences, people do. The "objective" interest is really a simplifying assumption in the tradition that treats states as actors in their own right. It is also fairly narrow because it only specifies what it takes to be the most important objective — security or power or wealth — and therefore may not provide much of a guidance when we want to deal with less apocalyptic issues. Still, there are many venerable schools of thought — which you will encounter in this course — that insist that we need not look below the abstract level of the state, or, if we do, we need not go very deep at all. Structural realism is among the former while classical realism, Marxism, and liberalism are among the latter.

The other two ways of looking at the national interest may be more helpful. Instead of postulating an objective to an abstract entity (the state), we take the national interest to be really an expression of individual preferences, whether they are elite decision-making groups or voters. In these views, a state implements the "best" policy consistent with either elite or voter preferences. The approaches tend to disagree as to who gets to decide what's "best" and whose preferences the policies will tend to reflect: those of the majority voters, of the few powerful members of the elite? However, they agree that somehow some relevant group of people has to agree on what the national interest is and how to get at it.

People have disagreements, usually vehement, on both of these issues. For example, you and I may disagree whether maintaining stable international markets is ultimately in our national interest. I, being internationally minded, may strongly believe that of America fails to keep the economy stable, it will eventually cause enormous problems domestically as well. You, being a firm agnostic about the value of globalization, may maintain that this is nonsense, and America should rely on its huge internal market and perhaps insulate itself as much as possible. There are many contentious issues in foreign policy, and what constitutes the national interest is a question that is seldom answered, although many talking heads seem to

assume that it is self-evident. Quickly: is preventing the spread of Islamic fundamentalism in the national interest? Or plugging the ozone hole? Or saving hundreds of thousands AIDS victims in Africa? Or assisting Israel against the Palestinians? Or the Palestinians against Israel? Or championing women's rights in Afghanistan? Which is more important? What about making sure Pakistan doesn't sell nuclear technology to other unsavory characters besides Lybia's ex-strongman Muammar Qaddafi? Or that Russia keeps its precocious bio warfare specialists from selling their services for hard currency? Or preventing Russia from gobbling up parts of neighboring countries? The list is potentially endless.

Suppose, however, that somehow we, as society, agree on what constitutes the national interest. For example, we all agree that America should strive to keep the global economy stable. We then fall into the next pit: what is the best way to do this? Should we maintain close links with repugnant regimes like the Saudi Arabia's autocrats just because they sit on the world's largest oil reserves that our European and Asian friends need so badly? Should we pursue a more hard-line policy in the Middle East to secure our ability to react to potential problems when the unpopular regimes eventually fall apart, as they must inevitably do? Or maybe we should hike up gas prices domestically so people don't drive needlessly? Or maybe we should invest heavily in fossil fuel-efficient technologies or even totally new hydrogen-based ones? Or perhaps tax the hell out of gas-guzzling SUVs that no sane person should be driving anyway? Or maybe everyone who thinks that Americans should be limited in their ability to drive tanks on highways is a goddamn pink Commie bastard that we should get rid of? New Yorkers and Bostonians with their nice public transportation and city lives that involve walking from place to place may be inclined to support policies that make driving costlier. But Californians and Texans who are rather spread out and who commute long distances may be much less enthusiastic. Anyway, even if we agree on the ultimate goal, we may still disagree ferociously on the methods we should use to get there.

Given all these disagreements that are bound to result from the simple fact that people are different, hold disparate beliefs, perceive the world in various ways, and have differential access to the levers of government, we should either appoint a dictator who simply implements the choices she wants (and hopefully these would be the ones she believes are for the good of the many) or else we must find a way to aggregate our disparate opinions into some sort of **collective choice**. (Note that even if we are ruled by a small elite, an oligarchy of sorts, then the members of this ruling elite must still find a way to aggregate *their* preferences into choices that the smaller collective body will make.) We have, however, already encountered a fundamental problem with group decision-making in the abstract setting above. Whether it is the elites or the voters who get to define the national interest or the means of achieving it, each group has to arrive at some ranking of alternatives and pick the one it likes best, and we now know that the institutional features of the group can be crucial in determining what group preferences will look like.

In other words, *it is impossible to conceive of the national interest solely in terms of preferences of the individuals that comprise the polity, even if these individuals are restricted to privileged elites. The national interest will depend on the institutional characteristics of the government, which themselves usually evolve after years of contentious politics, and thus tend to reflect the distribution of power in society.*

2.2 The Environment: Actions and Informational Structure

Actors do not make their choices in vacuum. The other defining component of our approach to international relations is the strategic environment in which interaction takes place. An environment is composed of **actions** that are available to the actors and an **information structure**.

The first is simply the set of actions which summarize how actors can interact. For example, during crisis negotiations, the set of actions might include (i) escalating the crisis by taking a provocative step, such as mobilizing troops or sending aircraft carriers into a volatile region, (ii) deescalating a crisis, (iii) starting a war, (iv) backing down and accepting the other side's demands, (v) producing new demands, (vi) insisting on previous demand and adopting a wait-and-see attitude, (vii) organize support of allies, (viii) make an offer on an unrelated issue linked to the opponent accepting your position on the one currently under consideration. The list can go on and on, although in most cases it is surprisingly short because it excludes all "irrelevant" choices. For example, although an actor may choose to produce more sugar, this choice will not be part of the crisis bargaining environment because it is not relevant for the decisions to be made in that strategic context. The environment limits the possible actions physically as well. For example, the action "initiate nuclear strike" is simply not available to non-nuclear powers.

The second component of the environment is its information structure. That is, what the actors can know and what they have to infer from observable behavior of others. This is related to beliefs because that information available in the environment determines in part the beliefs that the actors will hold. For example, suppose that in the crisis one side ostensibly deploys an armored division in an attempt to force the other to accept its demands. The move may appear aggressive, causing the other to update its beliefs and revise its estimate of the likelihood that its opponent is prepared to go to war. However, suppose that from its spies that side also learns that the tanks are old and there is insufficient fuel and supplies to actually put them in action. The deployment now appears as an empty bluff, and so the revised beliefs will very likely be different.

Thus, the actors (preferences and beliefs) interact in strategic environments (actions and information).

2.3 Strategic Interaction

Now, notice that I said "strategic" environment. What do I mean by **strategic interaction**? While we have defined the actors and the environment they operate in, we have not specified how outcomes are produced from their actions. The crucial aspect of interaction is that outcomes are not the result of any one actor's choices. Instead, in international relations, the choices of many actors determine outcomes.

An actor cannot choose an action simply because it has the best direct effect on the outcome it wants. Rather, it has to take into account the choices of others because they also affect the final outcome. So, an actor will choose an action both for the action's direct effect and its indirect effect on the actions of others. International politics is all about interdependent decision-making. That is, each actors does his best to further its goals knowing that the other actors are doing the same.

To give you a flavor of some of the issues involved, consider two social problems. The first, called the **Prisoner's Dilemma**, involves two actors who must decide whether they want to cooperate with each other or not. This game has four possible outcomes, they both cooperate, $\langle C, C \rangle$, only player 1 does, profileCD, only player 2 does, profileDC, and neither does, $\langle D, D \rangle$. Assume that each player's most preferred outcome is when only the other player cooperates, the second most preferred outcome is when both cooperate, the next to last outcome is when both defect, and the least preferred outcome is when he cooperates but the other player does not. For example, suppose the actors are states and "not cooperate" refers to implementing a protectionist economic policy (e.g., imposing a tariff on all goods imported from the other actor), whereas "cooperate" refers to maintaining free trade policies. Then, each player likes it best when it runs a protectionist policy itself (income from the tariffs and protecting competing domestic producers) but the opponent maintains an open regime (so the player's exports are sold on the opponent's country). Free trade is the next best regime, followed by a "tariff war" in which both countries impose tariffs that stifle trade. The worst outcome is to maintain an open regime while the opponent engages in protectionism.

	Protectionism	$\succ$	Free Trade	$\succ$	Tariff War	$\succ$	Open Policy
Country 1	(D, C)	$\succ$	(C, C)	$\succ$	(D, D)	$\succ$	(C, D)
Country 2	(C, D)	$\succ$	(C, C)	$\succ$	(D, D)	$\succ$	(D, C)
Payoffs	4	$>$	3	$>$	2	$>$	0

Table 2: Preferences in the Prisoner's Dilemma.

If you look at the preference orderings, you will see that 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 impose a tariff as well. Since free trade is the second-best outcome for both players, you might then conclude that this

should be the outcome produced by rational 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 dominant strategy in this scenario: it is always the best option for each player regardless of what the other player does. This means that the only rationalizable outcome is $\langle D, D \rangle$, the tariff war.

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 tariff war that will make both of them worse off relative to the free trade regime. They did not do this because they were stupid, irrational, or mistaken. They did this because their incentives in this situation are not aligned properly to support mutual cooperation.

You might be tempted to think that perhaps this outcome 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. Consider a game with two hunters who both most prefer to cooperate with each other but who might also be suspicious of each other's reliability. The problem, called the **Stag Hunt**, is due to Jean-Jacques Rousseau, and the story 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. You are one of these hunters. What do you do?

We set up the situation as a two-player game: you and the other hunter are the players. Each of you has two strategies: cooperate, C, or defect, D. There are four possible outcomes: both cooperate and catch the stag (Stag), you chase a rabbit and he stalks the stag (Tasty Rabbit for you, Hunger for him), you both hunt for rabbits (Stale Hare), and you stalk the stag while he catches a rabbit (Hunger for you, Tasty Rabbit for him). One possible specification of the payoffs that reflects the preferences is given in Table 3, which also rank orders the outcomes represented by the strategy profiles in which you are the first player.

Unreciprocated cooperation is the worst possible outcome for each player, and mutual defection is the second worst outcome. However, both players prefer mutual

	Stag	$\succ$	Tasty Rabbit	$\succ$	Stale Hare	$\succ$	Hunger
You	(C, C)	$\succ$	(D, C)	$\succ$	(D, D)	$\succ$	(C, D)
Other Hunter	(C, C)	$\succ$	(C, D)	$\succ$	(D, D)	$\succ$	(D, C)
Payoffs	4	$>$	3	$>$	2	$>$	0

Table 3: Preferences in the Stag Hunt.

cooperation to unilateral defection. Compare these preferences to the ones given in Table 2: 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. Figure 1 gives the full representation of this game that we are going to analyze. How would one play this game? The first thing to note is that which

		Other Hunter	
		C	D
You	C	4, 4	0, 3
	D	3, 0	2, 2

Figure 1: Payoffs in the Stag Hunt.

of *your* actions you prefer depends on what you think *your opponent's* action is going to be. If the other hunter is going to stalk the stag, you would get the stag if you cooperate as well, and you would get the juicy rabbit if you defect. Since the stag payoff of 3 is better than the juicy stag payoff of 2, your best response is to cooperate. If the other hunter is going after a rabbit, then trying to cooperate would just leave you hungry (with a payoff of 0), whereas chasing a rabbit would at least guarantee you a stale hare (payoff of 1).

Thus, in order to decide what you are going to do, you must predict what your opponent is going to do. The other hunter, however, faces a situation analogous to yours: *his* optimal action depends on what he thinks *you* are going to do. If he thinks you will cooperate, then he prefers to cooperate as well. If he thinks you will defect, then he 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.[6]

Consider now a situation where you cooperate but the other player defects: $\langle C, D \rangle$. If you expect the other player to defect, you will not want to cooperate either. But then the other player has no reason to expect you to cooperate, which means that we should not expect players to settle on this combination of strategies. An analo-

[6]This is called a Nash equilibrium.

gous argument applies to the case where you defect but your 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.[7]

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 hunters expects the other to cooperate, we end up with the cooperative outcome. If, on the other hand, we think of the hunters expects the other to defect, we end up with the non-cooperative outcome. So which expectation is more likely? Without knowing the hunters 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 hunters, 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 hunter?

In a sense, we are trying to protect ourselves from a mistaken expectation. Let's say I generally trust the other hunter to cooperate but I also know that sometimes he gets tempted when he sees rabbits, and I am not entirely sure that he will not see a rabbit or that if he sees one while stalking the stag, he won't abandon the stalking

[7] This is also a Nash equilibrium.

in order to chase after the rabbit. Now, if I cooperate, I would get the stag if he does not get distracted but I will end up hungry if he does. If I defect, I would get the juicy rabbit if he does not get distracted, and I will end up with a stale hare if he does. When I cooperate, the worst possible thing that can happen to me is to go hungry. When I defect, the worst possible thing that can happen to me is to end up with a stale hare. In that sense, defection is less risky because it leaves me less vulnerable in the case that I have misjudged my partner or he makes a mistake.

Since the other hunter can go through the analogous reasoning process, he will find that defection is less risky for him as well. More importantly, he will know that defection is less risky for *me*, and so might start thinking that this might make me more likely to choose to defect. Thus, he doubts about my willingness to cooperate will increase, and so his estimated risks of cooperating himself will go up. This will make defection even more tempting as the less risky strategy. Since I can reason all of this myself, I will conclude that the fact that I might be tempted to use the less risky strategy (defect) has made him more tempted to defect as well, which in turn increases *my* risks of cooperating, and so increases *my* temptation to defect. But this, of course, increases *his* temptation as well, and so on.

Even small initial doubts about the trustworthiness of one of the players can cascade in this interactive fashion and induce both players to choose the safe strategy of defecting, ending up with the outcome of mutual defection.

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.

What does this have to do with foreign policy? Consider two countries, both of which are considering building more armaments. Each can cooperate (not build) or defect (build). If both build, the military effects cancel each other out — so neither gets an advantage — but the build up itself is very expensive. This outcome is an *arms race*. If neither build, nobody gets an advantage as well, but nobody pays the extra costs either. This outcome is the *status quo*. Since the arms race does not alter the military balance relative to the status quo but does involve high costs, both players strictly prefer he status quo to the arms race. If one does not build but the other does, then the one that has failed to build is at a distinct disadvantage, and is forced to make political concessions. This outcome is *gain for the stronger player*, who prefers it to the arms race but since the buildup is so expensive, it cannot be offset by the gain, so he prefers the status quo to this unilateral advantage. The weaker player, on the other hand, likes these concessions the least of all possible outcomes.

You can verify that these preferences make the armaments game equivalent to a Stag Hunt. Our analysis of the latter can help rationalize a seemingly baffling outcome: an expensive arms race that gives neither side the advantage.

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 tariff war in a PD-like scenario is one of desire to exploit the other side's cooperative effort combine with a desire to avoid being saddled with the worst possible outcome. In this sense, the Stag Hunt is 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 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. 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.

When considering some particular interaction that these models seem to be appropriate for, you should think very carefully about the structure of the preferences. If you think the problem is analogous to 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 problem 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.

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!