Introduction to International Relations
Lecture 3: The Rational Actor Model

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

April 19, 2005

Overview. Continuing with our conceptual framework, we now study one of the most fundamental ideas: the rational pursuit of objectives. We look at what this means, what assumptions one has to make, and what problems one may have to deal with. We then examine the three fundamental features of the rational actor model: the actors, their environment, and the outcomes.

Outline of Lecture 3: The Rational Actor Model

1. Rational Actor Model

   a) (instrumental) rationality: purposeful behavior

   b) intelligence

   c) the (idealized) model

   d) optimal choice

      - mistakes
      - unintended consequences
      - deliberate risk

   e) utility and expected utility

   f) uncertainty

      - environmental
      - strategic

2. Critiques

   a) individual limitations (bounded rationality)

   b) organizational (principal-agent)

   c) social aggregation

3. Why model is still useful?

   - purposeful behavior despite mistakes
   - actors learn from experience
   - actors design institutions to overcome problems
   - analysis tools actively developed
   - what is the alternative?

4. Strategic choice

   a) actors: preferences, beliefs

   b) environment: actions, information

   c) strategic interaction

States do not act. People do. States do not make decisions. People do. States do not have goals. People do. Almost invariably, in order to analyze any sort of question in international politics, we would have to deal with individuals engaged in decision making. That is, individuals who are engaged in evaluating options, choosing among alternatives, and perhaps implementing their decisions. Ultimately, we would have to know how, given a particular environment with options, constraints, and information, people can analyze the alternative course of actions, and how they choose among them.

1 The Rational Actor Model

At the very basic level, we shall assume that people are **rational**. Rationality is a complex concept with different meanings to different people, so we shall try to make it precise what we mean by it. Rationality does *not* carry any connotations of normative behavior. That is, behaving rationally does not necessarily mean that one behaves morally or ethically. Hence, one may well argue that Hitler was rational even though his actions were clearly morally reprehensible. Similarly, it is quite possible to argue that suicide terrorists are rational even though most of us cannot relate to their choices. All we require is that the actions are somehow connected to the goals of the actor.

That is, rational behavior is **purposeful behavior**: an actor is behaving rationally if his choices are designed to achieve outcomes consistent with his goals. This is called **instrumental rationality** and is another way of saying that actors are able to relate means to ends, and they choose the means that help them obtain the ends they like most. Hence, if Hitler's goal was to exterminate the Jews, then the Final Solution was a rational act, and he was a rational actor (now, whether a sane person could actually have such a goal in mind is a different question entirely). If suicide terrorists believe that they will go to heaven (or that their families would be amply rewarded), then blowing themselves up becomes rational. That is, their acts, however incomprehensible to us, in fact do relate means to ends in a purposeful way.

The **rational actor model** treats foreign policy choices as products of the following idealized sequence. Given some problem, a rational decision maker takes into account the foreign policy *goals* of the nation and determines which ones take priority over others. Then, she identifies and analyzes the various *options* available. In her analysis, she traces the costs and benefits associated with each option, that is, she tries to estimate the likely *consequences* of making particular choices. This involves not just the gains and losses, but also estimating the relative likelihood of various outcomes. She then ranks the options from most preferred based on this analysis: on the bottom go options that are costly and unlikely to produce benefits close to the important goals, and on the top are options that are quite likely to work at no great cost. The decision maker then *chooses* the option that is ranked highest among the alternatives.

We shall call this option the **optimal choice**, and the prescribed course of action — optimal behavior. This is what we mean when we say that a person is *optimizing*. If you imagine that each outcome is associated with some number, called **utility** or **payoff**, then the decision maker is choosing the option that yields the highest payoff, so she is **maximizing utility**. When we talk about optimal choice, we do not assume that actors never make mistakes. In fact, we can evaluate what effect the possibility of making a mistake will have on overall behavior. Mistakes occur when actors erroneously evaluate the situation, either because they misunderstand it, or because they do not have the time or resources to do it properly. In addition to mistakes, actors' behavior may sometimes produce unintended consequences. That is, outcomes that the actors had failed to anticipate and that did not figure in the original evaluation of the option. We must distinguish between mistakes and unintended consequences on one hand, and deliberate risk on the other hand: sometimes actors may choose a course of action that may result in painful outcomes. If they do so deliberately and that outcome occurs, we cannot say that they have made a mistake: rather, they took a risk and the odds went against them. Taking risks can be a rational thing to do, and our model will accommodate that.

When we deal with choice under uncertainty, an action may produce one outcome with some probability and yet another with different probability. For example, if we decide to attack Iraq, then Saddam could have either capitulated immediately or fought to the death. Both (and many other possibilities) could occur, and we are not sure which one will, in fact, occur once the attack begins. However, we can form some beliefs and say that Saddam is unlikely to capitulate, so we may estimate the probability of outright capitulation to be low, say, 1%, and the probability of fighting to the death to be 99%. Suppose we have another option, which is targeted assassination: it could kill Saddam with probability 5% (and allow us to avert war) and it can fail with probability 95%, in which case we have to attack anyway but Saddam would have been warned that we're coming, so the war would be a bit more costly. Which option do you chose?

When we're dealing with uncertainty, we would calculate the **expected utility** from each option and then choose the one that yields the highest expected utility/payoff. In other words, we would behave as if we are **maximizing expected utility**. Roughly, this is a way to take into account the various probabilities attached to the outcomes and compute which ones are better than others.

There are two types of uncertainty that we have to deal with. **Environmental** uncertainty arises from the poor information actors have about the environment they are in, about other actors, or about the likely consequences of their actions. Because the real world is complicated and a variety of factors combine to produce outcomes, one can never predict with certainty what would happen if a course of action is chosen, how others will react to it, and what the eventual outcome will be. The situation is further worsened by other actors (opponents) actively trying to obfuscate your ability to predict. This **strategic uncertainty** arises from the rational behavior of the various actors and can seriously confuse matters. If your opponent knows that you will be trying to infer information from her actions, she may behave in a way specifically designed to prevent your from doing that. We shall see instances of this when we discuss crisis bargaining.

The branches of mathematics that deal with all these issues are called **decision theory** and **game theory**. Although we shall encounter some aspects of both, we shall not deal with them except peripherally in this course. It is sufficient to remember that these are mathematical tools for analyzing decision making in various environments. Both have made huge impact on the study of economics, political science, biology, and are making inroads in sociology and anthropology. You will probably see them from time to time as you continue your studies.

2 Cognitive and Social Critiques

To recapitulate, the rational actor model hypothesizes that people are rational in the sense that they choose actions that somehow help them achieve their goals.

This may not look controversial. However, one may well wonder about the ability of the actors to relate means to ends in an effective way. This is the **individual limitation critique**. For example, how much do the actors know about the environment in which they act? How capable are they of processing this information efficiently? Do they have enough time to ponder the alternatives? Are they operating under stress, or under mistaken assumptions? How limited are our cognitive capacities and how does this reflect on our rationality? These problems have to deal with the simple fact that we are imperfect humans with limited abilities who operate under uncertainty and under various constraints of time and resources. (In Chapter 7, the authors go in depth about various psychological biases and problems of bounded rationality. You should review these carefully.)

Yet another issue (not a problem because we can actually learn how to analyze it) with the model is that sometimes actors do not have control over the implementation of their decisions, the so-called **principal-agent problem**, or the **organizational critique**. This arises in situations where the decision maker, the principal, has to delegate to an agent the execution of the policy. In other words, you want certain goals achieved but your agent (who is usually better informed about the implementation details) is the one who actually has to do it. If your preferences are perfectly aligned with those of the agent, there is no problem. On the other hand, very often the preferences may be different, perhaps drastically different. The agent may simply sabotage your goals, or may implement the action in a way that thwarts other goals you have in mind. How to structure the incentives of the agent in such situations is the subject of a fairly large and sophisticated literature in economics. Hence, this is not so much of a problem as it is something that we have to carefully take into account.

Yet another problem with the model is that frequently various people participate in making some decision. Maybe it's a small committee empowered to analyze a problem and recommend solutions, or perhaps it is a huge unwieldy bureaucracy which produces policy outputs who knows how, or maybe it's a decision to be made by a referendum of the polity at large? It is not difficult to show that groups composed of entirely rational individuals can behave quite irrationally, this is the **social aggregation critique**. There are many psychological reasons for that (e.g. groupthink among others which are in your readings) but there are also even more fundamental problems. A very famous result due to Kenneth Arrow is the so-called *impossibility theorem*, which demonstrates that it is impossible to aggregate rational individual preferences in a way that guarantees that the resulting group (or social) preference would be rational itself. In other words, groups can sometimes behave unpredictably even though they are composed of entirely rational people.

These various problems with the rational actor model on the individual and group level may cause serious doubts in its usefulness as a tool for analysis. Yet such a conclusion would be premature. One must always keep in mind the limitations of the model, so much is true. However, there are many reasons why thinking about international politics in these terms is useful. First, people do act purposefully and even though they sometimes (often?) make mistakes, one should expect that generally their actions will reflect their goals. Second, people may not possess unlimited time and resources to collect all the necessary information, and they may not be able to optimize efficiently, but they can (and do) learn from experience. Any explanation that depends on people being perennially dumb, ignorant, and incapable of improvement should be immediately suspect. Third, the tool itself is being continuously developed to relax some of the more demanding assumptions that it makes. For example, now we can analyze optimal choice when the decision maker does not have complete and perfect information about the environment. We can analyze situations in which people make mistakes, or people do not have unlimited memories, or people do not have unbounded cognitive abilities. We can even analyze how people learn from repeated interaction and how they change their behavior. In other words, as the tool improves, our ability to get meaningful answers from its use will expand as well. Finally, one must ask oneself: if we do not use this rigorous model, then what should we use? What is the alternative? Unfortunately, whereas critics of the rational actor model are fond of jabbing at its weaknesses, they have failed to produce an alternative that is nearly as powerful. Until we have something better, we have to content ourselves with this model despite its shortcomings.

3 Strategic Choice

International relations is the study of the ways various actors interact internationally. It used to be defined very simply as the interaction among states, but even though nation-states are still the dominant way the international system is organized, many other entities actively participate in that interaction: international organizations (like the UN or the WTO), nongovernmental organizations (like Greenpeace, Amnesty International, or the Red Cross), bureaucracies and local governments, and even individuals. International relations is the study of the behavior of these actors; that is, the **interactions** among large organized groups of people. To analyze these interactions, we must distinguish three components: (i) the actors, (ii) the environment in which they act, and (iii) how outcomes are produced from the actions.

3.1 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 (EU), 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. As you should recall, we have referred to these as different **levels of analysis**.

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 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. Obviously, when we say "the U.S. is an actor," we are already deep into abstraction. The question is whether this particular one is useful (we shall see that sometimes it is, and sometimes it is not).

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.

3.2 The Environment: Actions and Informational Structure

Actors, of course, 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 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.

We shall deal with cases where actors do not possess perfect information. They may be quite unsure about what their opponents are doing and why. It is actually quite possible (although not very easy) to analyze these situations of incomplete information. We shall gain some intuition about what actors can do to attempt to elicit more information, and how they can make use of the sparse information that they already have.

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

3.3 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 just 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.

This is called “strategic interaction” and it can be extremely complicated because it involves forming expectations about what other actors are going to do, which in turn depends on what they think you are going to do, which, of course, depends on what you think they think you are going to do, and so on and so forth. Going through the chain of reasoning can be pretty difficult because you will end up in an infinite “I think that you think that I think that you think...” regression.

The tool for analysis of strategic interaction is called **game theory**, and it developed as a branch of applied mathematics early in the 20th century, but went nowhere until the US government financed researched for national security purposes in the mid 1960s. It was from these studies initiated for the purpose of finding ways of dealing with the Soviet Union that researchers discovered methods of dealing with uncertainty, beliefs, and strategic interaction in a productive way. In 1994, the Nobel prize in Economics went to three game theorists, the mathematician John Nash, the economist Reinhard Selten, and the strategic theorist John Harsanyi.