Robert Jervis, “Rational Deterrence: Theory and Evidence” (1989).

Jervis sketches the impediments to developing an expected utility model of deterrence. He then looks at biases in case selection and draws implications from the study of deterrence failure: Can these findings be explained by a rational theory? Deductive theory and expected utility models are not necessarily the same thing: Utility models can be applied to cases without attempting to generate new propositions. Deductive theory need not assume rationality on the part of the actors (eg. cognitive dissonance theory). Rational behavior can be idiosyncratic, and expected utility theories are not necessarily parsimonious. At the same time, prevalent irrationalities can produce strong, simple patterns.

Are Rational Deterrence Theories (RDTs) descriptive, prescriptive, or both? Many writings on RDT were meant to influence leaders, suggesting non-conformity with the theory's precepts. Case studies are usually descriptive and explanatory, but imply that different results could have been obtained had leaders behaved differently.

Debates and Distinctions.

Four main points are that (i) deductive theorizing and expected utility models are not necessarily the same, in that utility models can be simply be applied without proposing new relationships; (ii) expected utility arguments do not preclude idiosyncratic behavior, while prevalent *ir*rationalities can produce strong patters; (iii) there is no simple relationship between methodological and substantive arguments (eg. no inevitable connection between Subjective Expected Utility- SEU - and threat-based theories); and (iv) failed deterrence policy should not be equated with failure of deterrence *theory* (eg. deterrence successes do not confirm the theory if they occurred under conditions in which the theory predicted failure; and the failure of badly designed deterrents does not falsify the theory).

Rational Deterrence Theories.

Descriptive, prescriptive, or both? Does RDT deal with national behavior, international outcomes, or both? Greatest problem: Defining what RDT is. What sort of outcomes confirm RDT?

Empirical Criticisms and Second-Wave Deterrence Theory.

Three assumptions of Second-Wave DT are (i) adversaries are highly motivated to expand; (ii) adversary understands the state's defensive orientation; (iii) leaders on both sides are primarily concerned with the external situation. The latter assumption has been challenged by Lebow/Stein (deterrence policies are shaped according to domestic or personal interests and may thus not be in the national interest). Another potential error is to accept that states view the status quo as a future without war: If a state expects a significant deterioration in the SQ, it may engage in war with an expected utility of less than zero.

Second wave theorists emphasize ‘crisis instability:’ Deterrence might fail if a status quo power has insufficient strength, skill or resolve to resist, or an expansionist state draws incorrect conclusions about the weakness of the SQ power in these respects. Jervis points out that RDT is silent on how

leaders are to determine the intentions of their counterparts, and that it could be broadened to encompass "more tools of statecraft."

Bias in Case Selection.

General vs. immediate deterrence: In a case of immediate deterrence, the defender thought it had a defensible position and the challenger believed that it could succeed through coercion. However, the factors that go to the reconciliation of this conflict are likely to be very different from those that explain why it developed at all. Cases of successful general deterrence should therefore be examined: If we examine only cases in which one state wanted to challenge another, we ignore those cases in which this desire was successfully suppressed through deterrence.

How do we define 'failure'? Quantitatively? This is a red herring: "We may care less about *how often* deterrence fails than *how* it fails." Two reasons: (i) Strong versions of RDT are falsified if there are any cases in which the conditions for deterrence are met and yet it fails; (ii) in answering the "how" question, we are interested in *why* statesmen undertake such unsustainable steps as challenging credible, powerful commitments. The frequency of failure is irrelevant. Moreover, the focus on immediate deterrence ignores the important phenomenon of "paucity of superpower crises."

Generalization from Case Studies.

Case studies reveal limits on the ability to generalize: Strong, numerous idiosyncracies exist, and similar outcomes can be reached through different paths. Yet, structure and coherence are not entirely absent. Jervis identifies several generalizations, divided into the categories of (i) cognitive limitations (info processing); (ii) motivated biases; and (iii) the role of organizations in policy making.

(i) Cognitive Limitations:

1. A complex world demands "short-cuts" to rationality.
2. Beliefs are influenced by historical analogies to recent, important cases.
3. The role of accidents and confusion tends to be underestimated.
4. People often decide on the basis of a single, important value dimension.
5. What states think of themselves is often the reference point in assessing others.

(ii) Motivated Biases.

1. Statesmen tend to think well of their countries.
2. People tend to minimize perceived value trade-offs (pain avoidance principle).
3. Under bleak alternatives, statesmen exaggerate the chosen policy's chance of success.
4. Pressures to misperceive the environment as favorable vary directly with severity of loss.

(iii) Role of Domestic Politics and Org. Behavior in Implementation.

1. Domestic politics and coalition building explain part of expansionist policies.

2. Statesmen fail to realize discrepancies between policy making and implementation.

Implications:

1. Adversaries rarely understand each others' goals, fears, beliefs and perceptions.
2. The adversary will miss or misperceive many of the state's signals.
3. Objectively clear and credible commitments by one actor may not be perceived by another.
4. Actors overestimate the potency of threats and underestimate the utility of rewards/reassurance.
5. Threats and conciliation general need to be combined, but optimal mix and timing is difficult.
6. Successful immediate deterrence can weaken general deterrence (loser: increased grievances).

What cannot be explained by a Rational Theory.

Can SEU be applied if a policy fails perpetually because of incorrect supporting beliefs? Accepting outrageous beliefs will sacrifice common sense and utility. On the other hand, cognitive bias is consistent with rational behavior: Information and information processing represent costs, and people maximize utility when this is taken into account. One can thus argue that SEU is not disturbed by incorrect beliefs up to a certain threshold: "A full consideration of each bit of new information and the constant reconsideration of images and beliefs would be impossibly costly."

Problem (p. 200): This view assumes that people "somehow stop gathering and processing information just at the point at which the costs outweigh the expected gains. It is not clear how this point could be estimated – indeed trying to do so would itself be very costly." Moreover, cognitive biases should thus "operate much less"when important decisions are involved, but there is no evidence that this actually occurs. Thirdly, it is unclear that this approach "captures all the shortcuts to rationality, let alone anticipate ones not yet discovered" (*cf*. Rosen paper). Thus: Hard to see how SEU can succeed.

Does a rationality assumption cope better with motivated bias?

Motivated bias distorts the external environment to serve psychological or political functions. However, SEU may still operate: In instances of psychic pain avoidance (Nehru example) or absence of choice, motivated distortions may have only marginal or no influence on behavior. There may be "rationality in irrationality," in that people who consistently act in self-defeating ways may be driven by "subconscious ego dynamics," which is proximate to the idea that the costs of sacrificing goals is more than compensated by pain avoidance. Jervis acknowledges that this account of utilities is quite removed from classical RDT and only has deductive value if (i) discount rates are consistent and (ii) subjective costs of people's own "utility searches" are consistent.

Is Behavior Consistent?

SEU requires consistency in behavior and inferred preferences. Nations, however, do act inconsistently and thus an SEU explanation of behavior is impossible. The 'rational unitary actor' assumption is undermined by (i) coalitions which are held together by conflicting policy elements; (ii) deep internal disagreements which preclude decisional output; (iii) inconsistencies over time as

opposing factions come to power; (iv) cyclical majorities in policy development; and (v) inconsistent national policy due to non-integrated, divergent bureaucratic interests and perspectives.

People regulatory display inconsistencies that violate SEU (eg. consumer behavior): “Biased information processing, denigrating rejected courses of action, and bolstering the selected one not only make the person more comfortable with his decision, they also change the way he sees the world and the value he places on various outcomes.” SEU models will have to take better account of the coalition nature of politics and the operation of individual psychology in examining consistency in national behavior.