International Law and International Relations

Edited by Beth A. Simmons and Richard H. Steinberg

CAMBRIDGE www.cambridge.org/9780521861861

the distribution of peaceful nuclear technology to NNWS. Under the regime that eventually emerged, it was clear that the NNWS that joined the NPT would not suffer economic harm along this dimension.[45]

In addition to the concerns that played a major role at the review conferences, many other aspects of the uncertainty surrounding the distribution of gains from the NPT were largely or completely resolved during the initial trial period: In terms of security, the NPT greatly reduced the spread of nuclear weapons compared with what would likely have occurred without it. During the trial period, membership in the NPT increased to the point of being almost global.

In terms of the distribution of political gains (and losses), it became clear that concerns that the NPT would prevent European integration were groundless. What turned out to matter for European integration was not Britain's bombs but Germany's GNP. Time also rendered moot Japan's worries about its ability to react to a broad U.S. pullout from Asia.

In 1995 after four review conferences, the 163 parties to the treaty gathered in New York to decide whether the NPT would continue in force indefinitely or be extended for an additional fixed period or periods. Interviews with conference participants suggest that essentially all of the parties came to the conference favoring extension, a fact that itself provides powerful evidence of learning. Debate centered on whether extension would be indefinite or for a series of twenty-five-year periods.[46] In the end, a consensus resolution extended the NPT indefinitely. The NWS had gained what they expected to in terms of maintaining their power and influence (as Panofsky and Bunn note, "possession of nuclear weapons and permanent membership in the UN Security council remain identical"),[47] and the NNWS had learned how the NPT worked for them in practice.

* * *

[45] Nye presents additional examples of learning and uncertainty resolution in his discussion of policies relating to the nuclear fuel cycle and attempts to control aspects of the cycle related to nuclear weapons development. Nye 1981.

[46] The interviews appear in Welsh 1995. The debate regarding the extension provision was largely among the NNWS, since the NWS all favored indefinite extension. Ultimately, the Canadian argument that indefinite duration would cause the NWS "to be permanently held accountable to Article VI on disarmament" carried the day.

[47] Panofsky and Bunn 1994, 9.

CONCLUSION

The credibility of commitments in the face of uncertainty requires a trade-off between flexibility and constraint. I focus on a particular form of uncertainty – a one-shot uncertainty surrounding the division of gains from an agreement – and on a particular kind of flexibility – the combination of limited duration followed by a single renegotiation.

Some might argue that we do not observe much reneging on international agreements empirically and draw the implication that pretty much any duration and renegotiation provisions would do in a given context. In my model, this is clearly false. I calculate the lost utility from failure to choose the optimal provisions. In the case of small deviations, the lost utility takes the form of unrealized potential gains from the agreement. In the case of large deviations, it takes the form of reneging or failure to initiate the agreement at all. Empirically, a selection process similar to that described by George Downs, David Rocke, and Peter N. Barsoom regarding state compliance in international agreements is at work with respect to duration and renegotiation provisions.[48] The reason we do not observe much reneging in actual agreements is in part because their duration and renegotiation provisions have been chosen in ways that act to minimize this costly behavior. For example, the reason we do not observe agreements failing because of uncontrollable economic circumstances is that agreements in areas subject to such disruptions will tend to be of short duration – short enough that states experiencing sudden losses will stay with the agreement until it is renegotiated rather than renege.

My analysis also responds to some recent game-theoretic work in international relations. James D. Fearon points out a weakness in current theories of international cooperation that focus primarily on the enforcement of international agreements while ignoring the bargaining that generates the agreements in the first place. Fearon's model integrates the bargaining over the terms of an agreement into the cooperation problem. This formulation reveals that the same shadow of the future that allows self-enforcing agreements also makes reaching an agreement more difficult by increasing the distributional effects of the selection of the initial equilibrium.[49]

[48] Downs, Rocke, and Barsoom 1996.

[49] Fearon 1998. Fearon's work as well as Morrow's show how distributional differences can undermine cooperation in significant ways. Morrow 1994. These works are in part a response to Grieco and Krasner, who have rightly argued that neoliberals tend to emphasize enforcement issues and ignore distributional issues. See Grieco 1988; and Krasner 1991.

I provide a theoretical justification for the fact that states often integrate planned renegotiation into international agreements. Building in renegotiation at the start reduces the distributional impacts of the initial equilibrium selection but does not reduce the shadow of the future that supports enforcement of the agreement as a whole since violations in the initial period can still be punished by noncooperation in future periods. Of course, allowing renegotiation adds the possibility that one party may drop out at the end of a particular finite-duration contract. Indeed, in the case of the International Coffee Agreement, the United States did just that. Nonetheless, if the probability of continuation is sufficiently high, finite-duration contracts linked by renegotiation may represent the real-world solution to Fearon's theoretical dilemma.

Finally, my analysis extends the neoliberal international relations literature beyond its current focus on the general issue of how cooperation can emerge. Both Robert Axelrod and Kenneth Oye suggest devices such as lengthening the shadow of the future, practicing reciprocity, and improving recognition capabilities; Stephen D. Krasner looks at the role of international regimes in promoting and maintaining cooperation; and Robert O. Keohane argues that regimes reduce the transactions costs associated with international cooperation.[50] This literature has opened up the central questions of international politics. It has done so, however, only by moving well away from any detailed analysis of specific institutional arrangements or questions of institutional design. In other words, this literature has failed to investigate the precise mechanisms through which cooperation can emerge.[51] There is no inherent reason, however, why the broader political issues cannot be considered simultaneously with the specific institutional arrangements designed to address them in ways that illuminate both the broader relationships and the institutionalization itself. My goal in investigating duration and renegotiation provisions has been to deepen our understanding of international cooperation by asking about specifics.

[50] See Axelrod 1984; Oye 1986; Krasner 1983; and Keohane 1984.

[51] Likewise, the tools of game theory have been directed mainly at abstract questions that emphasize cooperation rather than institutional design as the dependent variable.

16

Driving with the Rearview Mirror: On the Rational Science of Institutional Design

Alexander Wendt

How can social scientists best contribute to the design of international institutions? Presumably our value lies in producing knowledge about design that those designing institutions need but do not have. But what kind of knowledge is that? What should a science of institutional design be "about?"

As a discipline international relations (IR) has barely begun to think about institutional design. Anarchy makes the international system among the least hospitable of all social systems to institutional solutions to problems, encouraging actors to rely on power and interest instead. *** Skeptics may be right that all this activity is unimportant but policymakers apparently disagree. And that in turn has left IR with less to say to them than it might have. By bracketing whether institutions matter and turning to the problem of institutional design, therefore, this volume takes an important step toward a more policy-relevant discourse about international politics.

The articles in this volume deserve to be assessed on their own terms, within the particular rationalist framework laid out in Barbara Koremenos, Charles Lipson, and Duncan Snidal's introduction. That framework highlights collective-action problems and incomplete information as impediments to institutional design. *** However, offering an internal critique of the Rational Design project from any rationalist perspective is not something I am particularly qualified or inclined to do, nor was it the charge given to me when I was generously invited to contribute. From the start

For their helpful comments on a draft of this article, I am grateful to two anonymous reviewers, the *IO* editors, Michael Barnett, Deborah Boucoyannis, Martha Finnemore, Peter Katzenstein, and especially Jennifer Mitzen.

the editors deliberately set aside a number of "nonrationalist" arguments in order to see how far they could push their approach to the problem. The purpose of soliciting this comment was to get an outside perspective.

Actually, I am not that qualified or inclined to make a fully external critique either. Although some epistemological issues will come up, I share the volume's commitment to social science, and while I doubt that rationalism can tell us everything, I certainly think it can tell us a lot.[1] Additional insights about institutional design might emerge by rejecting social science or rationalism altogether, but I shall not do so here. However, in the space between a purely internal and purely external critique I hope to raise some fairly fundamental questions about the approach. ***

I shall raise two main concerns, one more external than the other. The first is the volume's neglect of alternatives to its explanation of institutional designs. At base, the theory of rational design is that states and other actors choose international institutions to further their own interests.[2] This amounts to a functionalist claim: actors choose institutions because they expect them to have a positive function.[3] Alternatives to this hypothesis come in at least two forms, both associated with "sociological" or "constructivist" approaches to institutions.[4]

On the one hand, alternatives could be rival explanations, where the relationship to the theory of rational design is zero-sum; variance explained by one is variance not explained by the other. At first glance it might seem hard to identify plausible rivals. One is tempted to say, Of course actors design institutions to further their interests – what else would they do? But in fact there are some interesting rivals, both to the proposition that institutions are rationally chosen and to the proposition that they are designed. I discuss each in turn and argue that neglect of these alternatives makes it more difficult to assess the volume's conclusions. ***

On the other hand, "alternatives" could refer to explanations that do not contradict rational-design theory but embed it within broader social or historical contexts that construct its elements (preferences, beliefs, and so on). Whereas the question with rival explanations is one of variance explained, the issue here is one of "causal depth."[5] Even if states choose rationally, this may be less interesting than the underlying structures that

[1] See Wendt 1999; and Fearon and Wendt forthcoming.
[2] Koremenos, Lipson, and Snidal [2001], 762.
[3] On functionalism in design theory and its alternatives, see especially Pierson 2000b.
[4] For good introductions to this extensive literature, see Powell and DiMaggio 1991; Hall and Taylor 1996; and March and Olsen 1998.
[5] Wilson 1994.

make certain choices rational in the first place. It is on such structures that sociological and constructivist approaches to institutions typically focus. ***

* * *

Despite its focus on alternative explanations, this first critique remains internal in the sense that it assumes, with the Rational Design project, that the question we are trying to answer about institutional design is an explanatory one: Why do institutions have the features they do? However, part of what makes the problem of institutional design interesting, in my view, is that it raises further questions which go beyond that explanatory concern. In particular, the term *design* readily calls up the policy-relevant question, What kind of institutions should we design? *** [Given] that this volume focuses on a theoretic issue with important policy implications, it seems useful in this essay to reflect on how the gap between positive and normative could be narrowed further.

Bridging this gap depends, I shall argue, on recognizing the epistemological differences between the kinds of knowledge sought in the scientific and policy domains, which stem from different attitudes toward time. Positive social scientists are after "explanatory" knowledge, knowledge about why things happen. This is necessarily backward-looking, since we can only explain what has already occurred ***. Policymakers, and institutional designers, in contrast, need "making" or "practical" knowledge, knowledge about what to do. This is necessarily forward-looking, since it is about how we should act in the future. As Henry Jackman puts it, "we live forwards but understand backwards."[6] The former cannot be reduced to the latter. Knowing why we acted in the past can teach us valuable lessons, but unless the social universe is deterministic, the past is only contingently related to the future. Whether actors preserve an existing institution like state sovereignty or design a new one like the EU is up to them. ***

Practical knowledge may nevertheless interact in interesting ways with explanatory knowledge. To show this, in the last third of this article I briefly discuss two domains of inquiry about institutional design not addressed in this volume. The first is institutional effectiveness. *** The second domain is the specifically normative one. What values should we pursue in institutions? ***

Positive and normative inquiries are, of course, in many ways distinct, but a science of institutional design that deals only with the former

[6] Jackman 1999.

will be incomplete and useful primarily for "driving with the rearview mirror." The larger question I want to raise here, therefore, is an epistemological one – what should count as "knowledge" about institutional design? In social science we often assume that knowledge is only about explaining the past. Institutional design is an issue where the nature of the problem – making things in the future – may require a broader view ***.

ALTERNATIVES TO RATIONAL DESIGN

Given the question, What explains variation in institutional design? it is clear that rational-design theory provides some leverage. But how much leverage? It is difficult to say until we make lateral comparisons to its rivals and vertical comparisons to deeper explanations. Thus, assuming that the phrase "rational design" is not redundant, I break the volume's hypothesis down into two parts, that institutions are chosen rationally and that they are designed.

Alternatives to "Rational"

*** Rationality can be defined in various ways.[7] In rational-choice theory it refers to instrumental or "logic of consequences" thinking:[8] Actors are rational when they choose strategies that they believe will have the optimal consequences given their interests. *** This is a subjective definition of rationality in that a rational choice is not what will actually maximize an actor's pay-offs (we might call this an "objective" view of rationality), but what the actor thinks will do so. ***

If for a single actor rational action is what subjectively maximizes its interests, then when there are multiple actors, as in international politics, a rationally chosen institution will be one that solves their collective-action problem ***. *** Collective-action problems, in short, are subjective at the group level, in that they are constituted by a shared perception of some facts in the world as (1) being a "problem" (versus not), (2) requiring "collective action" (versus not), and (3) having certain features that constitute what kind of collective-action problem it is (coordination, cooperation, security, economic, and so on). These understandings are only partly determined by objective facts in the world ***. They are also constructed by a communicative process of

[7] See especially Hargreaves-Heap 1989.
[8] Jackman 1999.

interpreting what that world means and how and why designers should care about it.[9] ***

* * *

What are the alternatives to the hypothesis that states choose subjectively rational institutions? One, of course, is that states knowingly choose institutions that will defeat their purposes, but that does not seem very plausible. We have to look elsewhere for interesting alternatives. I discuss two.

The Logic of Appropriateness

One alternative is that states choose institutional designs according to the "logic of appropriateness":[10] Instead of weighing costs and benefits, they choose on the basis of what is normatively appropriate. *** In international politics there are many examples of decision making on appropriateness grounds. An example I have used before is what stops the United States from conquering the Bahamas, instrumental factors or a belief that this would be wrong?[11] One can construct an "as if," cost-benefit story to explain nonconquest, but I doubt this is the operative mechanism; it is more likely that U.S. policymakers see this as illegitimate. A more difficult and thus interesting example is provided by Nina Tannenwald's study of the "nuclear taboo," which suggests that even when instrumental factors weighed in favor of using nuclear weapons, as in the Vietnam War, U.S. decision makers refrained on normative grounds.[12] The way such a logic ultimately works is through the internalization of norms. As actors become socialized to norms, they make them part of their identity, and that identity in turn creates a collective interest in norms as ends in themselves.[13] The result is internalized self-restraint: actors follow norms not because it is in their self-interest, but because it is the right thing to do in their society. ***

The Bahamas and nuclear taboo examples highlight the fact that the logic of appropriateness has usually been used in IR to explain *compliance* with regimes.[14] *** However, design is a different question from compliance, to which it is less obvious that logics of appropriateness are directly relevant.

[9] Kratochwil 1989.
[10] March and Olsen 1998.
[11] Wendt 1999, 289–90.
[12] Tannenwald 1999.
[13] Wendt 1999.
[14] The Meyer School being an important exception.

Nevertheless, there are at least three ways in which normative logics might be rivals to rational explanations of institutional design. One is by supplying desiderata for institutions that make little sense on consequentialist grounds. A norm of universal membership, for example, operates in many international regimes. Why do landlocked states have a say in the Law of the Sea, or Luxembourg a vote in the EU? It is not obvious that the answers lie in the enforcement and distributional considerations emphasized by the Rational Design framework. Or consider the norm that Great Powers have special prerogatives. Without reference to this idea, it is hard to explain the inclusion of Russia in the Group of Eight, or to make sense of debates about the future of the UN Security Council. The norm that the control of international institutions should be democratic is also gaining strength. The Rational Design framework proposes that designs for institutional control reflect degrees of uncertainty and asymmetries of contribution, yet in debates about how to fix the "democratic deficit" in the EU and other international organizations such cost-benefit considerations seem less salient than questions of legitimacy and principle.[15] *** And so on. These possibilities do not mean that rational factors are not also operative in regime design, but they do suggest the story may be more complicated than a pure consequentialism would allow.

A second, converse, way in which logics of appropriateness may constitute rival hypotheses is by taking design options that might be instrumentally attractive off the table as "normative prohibitions."[16] *** [One] might expect a purely rational regime for dealing with "failed states" to include a trusteeship option, but because of its association with colonialism, this is unacceptable to the international community. Finally, norms about what kinds of coercion may be used in different contexts may also factor into regime design. Military intervention to collect sovereign debts was legitimate in the nineteenth century,[17] but it is hard to imagine this being done today. *** A true test of rational-design theory would include *all* instrumentally relevant options, not just those that are normatively acceptable.

Finally, logics of appropriateness can affect the modalities used to design institutions, which as a result may be historically specific. ***

In at least three ways, then, logics of appropriateness may help structure international institutions. These possibilities do not mean that

[15] See, for example, Pogge 1997; and Dryzek 1999.
[16] Nadalmann 1990.
[17] Krasner 1999.

consequentialism is wholly absent. But insofar as our objective is to assess variance explained, the logic of appropriateness suggests that rival factors may be important as well.

* * *

On Uncertainty

In addition to instrumental thinking, rationality as understood in this volume relies on a particular, and contested, way of handling uncertainty. As the editors point out, a focus on uncertainty is one of the Rational Design project's significant departures from earlier rationalist (and non-rationalist) scholarship on international institutions.[18] Since uncertainty is intrinsic to social life, and especially to institutional design – which tries to structure an otherwise open future – addressing it can make IR more realistic and policy-relevant. However, the Rational Design framework seems to treat the nature of uncertainty as unproblematic and ends up with a conceptualization that effectively reduces it to risk. This assertion may seem wrong, since the editors say they are adopting the "standard terminology in using the term *uncertainty* instead of *risk*," [19] but the premise of this terminology is that the two are equivalent. That there is an important distinction between risk and uncertainty has been known at least since Frank Knight's classic 1921 work[20] and the distinction is used in some rationalist scholarship today, even elsewhere by Snidal himself.[21] But in most orthodox economics and formal theory the two are conflated, and it is to this literature that this volume seems most indebted. In contrast, heterodox Austrian and post-Keynesian economists vigorously uphold Knight's distinction and indeed base much of their critique of mainstream economics on its failure to do so.[22] ***

"Risk" describes a situation in which some parameters of the decision problem, such as other actors' preferences or beliefs, are not known for certain, but – importantly – all the possibilities are known and can be assigned probabilities that add up to 1. The utility of different courses of action is then weighted by these probabilities, leading to the formalism of expected-utility theory. A key implication of risk is that even though actors

[18] Also see Koremenos 2001.
[19] Koremenos, Lipson, and Snidal 2001, 779.
[20] Knight 1921.
[21] For example, Abbott and Snidal 2000, 442.
[22] The literature here is extensive. See, for example, Davidson 1991; Vercelli 1995; and Dequech 1997.

cannot be certain about the outcomes of their choices, they can at least see well-defined (if still probabilistic) relationships between ends and means, so that they can calculate precisely the chances of achieving their goals with different strategies.[23] Choose *A*, and there is a given chance that payoff *X* will occur; choose *B*, another chance; and so on. This is significant because it means there is always a clear and principled answer to the question, What is the rational thing to do?

*** [Uncertainty] exists when an actor does not know all the possibilities in a situation, cannot assign probabilities to them,[24] or those probabilities do not sum to unity. To distinguish it from the standard view, uncertainty in this heterodox tradition is often qualified with adjectives like "strong," "hard," "genuine," or "structural." *** [Where] there is genuine uncertainty, the clear (if probabilistic) relationship between ends and means breaks down, so that optimal behavior may not be distinguishable from sub-optimal. If optimality is no longer calculable, then what is instrumentally rational is no longer well defined.

This suggests a rival hypothesis about how rational actors should behave. On the orthodox view, actors facing incomplete information should continually adjust their beliefs and strategies in response to changing estimates of the situation. The importance of such updating is reflected in the volume's conjectures about the effects of uncertainty on rational design, namely that institutions should maximize flexibility and individual control. In contrast, Ronald Heiner argues on heterodox grounds that actors facing genuine uncertainty may be better off *not* trying to optimize, because they are not competent to grasp the true problem and so are prone to make mistakes and have regrets.[25] On his view, in other words, in situations of genuine uncertainty expected-utility theory may actually be a poor guide to "rational" behavior. Instead, actors should do just the opposite of what that theory recommends: follow simple, rigid rules and avoid continually updating expected values. Heiner argues further that most people in the real world understand this, since their behavior is much more stable than would be expected if they were constantly optimizing. Under conditions of genuine uncertainty, it is our willingness to *depart* from the optimizing standard that is the "origin of predictable behavior."[26] In the context of institutional design, therefore, the rational

[23] Beckert 1996, 819.
[24] Which may presuppose a nonsubjectivist view of probability.
[25] Heiner 1983.
[26] Ibid.