Formal Models of International Institutions

Michael J. Gilligan[1] and Leslie Johns[2]

[1]Department of Politics, New York University, New York, New York 10003; email: michael.gilligan@nyu.edu

[2]Department of Political Science, University of California, Los Angeles 90095; email: ljohns@polisci.ucla.edu

Annu. Rev. Polit. Sci. 2012. 15:7.1–7.23

The *Annual Review of Political Science* is online at polisci.annualreviews.org

This article's doi: 10.1146/annurev-polisci-043010-095828

1094-2939/12/0615-0001$20.00

Keywords

Abstract

The past three decades have witnessed the development of a rich literature that applies the formal tools of game theory to understanding international cooperation and international institutions. We divide this literature into three "generations" of scholarship. With a few notable exceptions, the first generation used very simple models—2 × 2 normal form games—to understand why states need to cooperate and why they comply with their cooperative agreements under conditions of anarchy. This first generation unfortunately bogged down in the neorealist–neoliberal debate. Second-generation scholars began to use tailor-made models to address the neorealist–neoliberal debate and to turn to new questions, such as how international agreements are created and how domestic political divisions affect international cooperation. With answers to the key questions of how international agreements are created and complied with, third-generation scholars could turn to increasingly refined models to answer specific questions about international institutions, such as the proper size of multilateral agreements, how the gains of cooperation are distributed, whether flexibility provisions should be built into agreements, and the specific functions of international organizations.

INTRODUCTION

Over the past three decades, international relations (IR) scholars have developed a rich literature using the tools of game theory to understand international cooperation and institutions. We divide this research into three generations, although the borderlines between them are somewhat arbitrary. The first generation, which arose in the mid-1980s, primarily addressed why states create and comply with international agreements when there is no international state to enforce them. This literature relied heavily on the repeated prisoners' dilemma, reputation, and hegemonic stability arguments.

This first generation said little about how international agreements are created except in the special case where a hegemon creates them. First-generation models also mainly assumed that states are unitary actors and did not address the effect of domestic politics on international cooperation. This first generation became mired in a neorealist–neoliberal debate, which amounted to an irreconcilable disagreement over assumptions about states' utility functions.

The second generation, which arose in the mid-1990s, addressed these three shortcomings. It applied formal models of bargaining to examine how states negotiate international agreements in the absence of a hegemon. These models led to new connections between the bargaining over an agreement and subsequent compliance. Several second-generation models included domestic politics, which generated new insights into why states comply with their agreements and how states bargain to create them. However, the most sophisticated of these models failed to generate testable predictions. The neorealist–neoliberal debate disappeared in the second generation, but not before some scholars made important points about the role of formal models in the social sciences and showed that the two camps could be reconciled with a proper approach to modeling.

Third-generation scholars moved on to more refined issues of international cooperation and institutions, such as distribution, depth, flexibility, multilateralism, and the functions of international institutions. Third-generation research explored the following questions:

- *Distribution*: Which countries gain more from international negotiations and why?
- *Depth*: Are deeper agreements more effective? Does compliance with a treaty mean that it is effective?
- *Flexibility:* When should rules be rigid and when should they be flexible to allow some cheating?
- *Multilateralism:* How does the size of the membership of an agreement affect the depth of cooperation?
- *Functions of international institutions:* How and why are institutions used to authorize the use of force, manipulate domestic politics, develop bureaucratic expertise, and adjudicate disputes?

By necessity we exclude several topics. We do not examine the literatures on international mediation and alliances because they study the prevention and management of security crises and crisis bargaining, not the creation of cooperative surpluses. Kydd (2010) provides an excellent overview of mediation models. We do not discuss the European Union because it has become more like a federal organization than an international one. Scholars may debate whether the EU is an international or a supranational institution, but it is qualitatively different from the types of institutions we discuss here.

FIRST-GENERATION MODELS

First-generation models focused on why states comply with their cooperative agreements under anarchy. A special issue of *World Politics* did much to popularize this topic (Oye 1985). This

literature argued that states' policies can generate inefficient externalities. For example, one state's trade barrier can reduce the welfare of countries that export to that state more than it increases the welfare of the state that enacts the barrier. Similarly, in the realm of environmental politics, states may not take into account the costs of their pollution for their neighbors. They would therefore pollute beyond the point where the marginal benefit of the pollution to themselves (from production processes) is equal to the marginal cost to themselves *and* neighboring states. This generation mainly conceptualized international cooperation problems as 2×2 normal form games, including the prisoners' dilemma and occasionally stag hunt and coordination games (Wagner 1983).

Bilateral Cooperation: The Repeated Prisoners' Dilemma

One way to provide micro-foundations for the prisoners' dilemma (PD) is to assume that two states must each choose a policy (such as a tariff or a level of environmental regulation) that affects the welfare of the other. Suppose each country chooses its policy by maximizing only its own utility. Each state will set its own marginal cost of the policy equal to its own marginal benefit, ignoring the impact of the policy on the other state. This creates a deadweight loss. If each country takes into account the effect of its policy on the other, this creates a cooperative surplus. If the payoffs from the status quo are Q and the payoffs from cooperating are C, then $C > Q$.

There is a temptation to cheat (defect) on the agreement though. By promising to cooperate but then reverting to a noncooperative policy, a state can maximize its utility with respect to its own policy and still benefit from the other state's cooperation. Cheating while the other country cooperates offers even greater utility than mutual cooperation. If the payoffs from these cheating or "temptation" policies are T, then $T > C$. For the state that is cheated, however, the payoff is the worst possible. This country suffers losses from its own policy (because it does not set its marginal benefit equal to its own marginal costs) but is not compensated by the other state's cooperation. If this "sucker" payoff is S, then $Q > S$.

We now have a preference ordering over all possible outcomes, $T > C > Q > S$, that generates the familiar PD. The only equilibrium of this one-shot game is mutual cheating. With no world-wide state to enforce cooperative agreements, states should not comply with them. Understanding why states would nonetheless cooperate under anarchy was the central focus of first-generation scholars.

One answer is repetition. If states interact with each other an infinite number of times, then they can use so-called "trigger strategies." The simplest of these is a bilateral trigger between two countries. Suppose state A sets its policy at the cooperative level as long as state B cooperates, but reverts to the status quo policy if state B cheats. This implicit threat is credible because the status quo policies are a Nash equilibrium. Cheating on the agreement triggers a reversion to the noncooperative outcome for some number of periods. This number of reversion periods can be any value from one to infinity. The latter is often called the "grim trigger." If states' discount rates are sufficiently high—or, as Axelrod (1985) put it, if "the shadow of the future" looms sufficiently large—they will comply with the agreement. Call the two states' discount rates δ. Assuming the grim trigger, states will comply with the cooperative agreement if

$$\frac{C}{1-\delta} \geq T + \frac{\delta Q}{1-\delta} \Leftrightarrow \delta \geq \frac{T-C}{T-Q}$$

The importance of repetition was popularized by Axelrod's *Evolution of Cooperation*. Axelrod did not write a formal model of international cooperation. Instead, in an early example of what we now call agent-based modeling, he ran a tournament in which entrants submitted competing

strategies for a repeated PD. He created a computer program that pitted these strategies against each other to see which performed the best over time. Though not a formal model, Axelrod's work emphasized the importance of repetition in enforcing cooperation under anarchy.

Multilateral Cooperation: Hegemons and Law Merchants

Multilateral cooperation is more complicated. The word multilateral has normative connotations for many IR scholars (Ruggie 1992), but for our purposes, multilateral cooperation simply means cooperation between more than two countries. The first problem of multilateral cooperation is "free riding" when the benefits of cooperation are not excludable. Imagine five countries whose pollution affects each other's welfare. If countries A and B cooperatively reduce their pollution, then countries C, D, and E enjoy the benefits of lower pollution even though they do not abate their own. A second problem is monitoring countries' behavior. In the bilateral case, if country A cheats, then B should know this, or at least be able to determine the likelihood that cheating occurred. In a multilateral setting, however, if country A cheats on B, how are C, D, and E to know? Country B can accuse country A of cheating, but how are C, D, and E to know that this accusation is justified? The problem of free riding was addressed in the first-generation literature by hegemonic stability theory (HST). The problem of monitoring was examined by an institutional economics literature on medieval trading institutions (Milgrom et al. 1990).

Little of the hegemonic stability literature can be characterized as formal modeling. However, HST addressed the central question of first-generation formal modelers: why do states comply with international agreements? HST, which is often traced to Kindleberger (1986), begins with the premise that the provision and enforcement of international cooperative arrangements is a public good and, therefore, prone to undersupply. Most countries have incentive to free ride by failing to comply with and enforce agreements, but a hegemon does not. HST claimed that a hegemon would supply the public good because its size and connections with other countries made it a high demander of international stability. HST also argued that hegemons tend to decline. This creates an interim period of instability in which neither the old nor the new hegemon supplies the public good. These claims were not the conclusions of a well-developed model but were conjectures based on one or two cases (Gilpin 1981, Kennedy 1987).

A charitable reading of Kindleberger suggests that he was not making a causal argument but rather a policy argument in favor of continued U.S. leadership in the world economy following the collapse of the Bretton Woods system. Still, the theory provoked a large response in the IR literature. Snidal (1985) offered an analysis with some formal elements. He argued that even if agreements and enforcement are public goods, this does not imply that small states will not supply them. He showed that small states may take over provision of the public good after hegemonic decline.

One major HST model was by Alt et al. (1988). This is an incomplete-information model that explains hegemonic decline. Potential free riders decide in each period whether to comply or cheat. The hegemon then decides whether or not to punish defections. Punishing is costly for the hegemon. The size of these costs is the hegemon's private information. When potential free riders decide whether to defect, they have incomplete information about the hegemon's costs of punishment. In the most interesting case, would-be free riders and the hegemon play mixed strategies, sometimes defecting and sometimes punishing, respectively. Hegemonic decline is an increase in the number of periods of unpunished defection. Alt et al. explain this decline as a result of belief updating. As free riders observe the hegemon's punishment behavior, they get a clearer picture of how the hegemon's costs of punishment are distributed and then increasingly challenge the hegemon.

Turning to the monitoring issue, Milgrom et al. (1990) showed that institutions can help enforce multilateral cooperation by sharing information about past defections. In their model, a group of traders want to promote cooperative behavior. However, interactions between any two members are too rare for the bilateral repeated PD to be an effective enforcement strategy. Instead, the members agree that they all will punish (boycott) a cheater even if they were not cheated themselves. Thus, a cheater will be punished even if he never again interacts with the party he cheated. This solution raises a new problem, however. How can these third parties know whether cheating has occurred? According to Milgrom et al. (1990), this problem was solved in medieval trade by the creation of a "law merchant" who adjudicated disputes and shared information about past interactions. Third parties could coordinate their actions on the law merchant's judgment.

This work made two important contributions. First, it showed that institutions can solve the problem of monitoring compliance in large communities. Second, it showed how voluntary punishment could be individually rational even if the punisher was not the party that was cheated. More than two decades later, similar models are used to understand international courts and dispute resolution procedures.

One simplifying assumption of the law merchant model is that cooperation is a "club good," not a public good. It is nonrivalrous in consumption but excludable because merchants can exclude cheaters from the benefits of cooperation by refusing to trade with them. Such excludability is not always possible in international cooperation. For example, suppose a hegemon must enforce a pollution treaty. If other targeted punishment policies, such as the withdrawal of foreign aid or trade concessions, are not available, then the hegemon can only punish the cheater by raising its own level of pollution. This punishes compliant countries as well. This highlights one of the appeals of HST for policy-oriented scholars like Kindleberger: a hegemon could offer other states a bundle of policies and inducements, some of which were targetable and some of which were not. These targetable policies and inducements (such as foreign aid or trade concessions) could be used to enforce cooperation with the nontargetable policies without imposing collateral damage on other compliant states.

The Neorealist–Neoliberal Debate

The harshest early critics of this first generation were neorealists, who dubbed the nascent theory of international cooperation "neoliberal institutionalism" (Grieco 1988). Neorealists agreed that compliance under anarchy is difficult because states want to avoid the sucker payoff. However, neorealists believed that states also fear that their partner may use cooperative gains to attack them in the future. The most important feature of anarchy for neorealists is that states are responsible for their own security. Today's cooperative partner may be tomorrow's enemy. Therefore, states must maximize their gains from cooperation relative to those of other countries.

The neorealist–neoliberal debate produced more heat than light, but it offered an opportunity to teach IR scholars about the usefulness of formal modeling. Powell (1994) argued that the neorealist–neoliberal debate was unproductive because it stemmed from a misunderstanding about the purpose of models. Rather than creating a model in which states are so concerned about future attack that they will not cooperate with each other, realists simply assumed that this was so. By assuming that countries maximize relative gains (rather than modeling a situation that induces this concern), the realists did not add anything to their intuition that cooperation is harder than the neoliberals think it is. Powell (1994) showed that countries' concern with relative gains is conditional on other factors that determine the likelihood of war. By doing the modeling correctly, we actually learn why or how anarchy hinders cooperation rather than just asserting it.

Before closing our discussion of first-generation models, we must mention the important work of Robert Keohane. Though not a formal modeler, Keohane (1984) derived keen insights about the possibilities and limitations of international cooperation from the Coase theorem. In a Coasian world, states should have no problem creating international agreements if the gains from cooperation are greater than the transaction costs. The puzzle then : why don't states always cooperate? Keohane recognized that the requirements of the Coase theorem—low transaction costs and well-defined property rights—were absent in international politics. Institutions were more functional—they reduced transactions costs so as to make Coasian bargains possible (Gilligan 2009).

First-generation models are relatively simple, yet they answer the main question of scholars at that time: why do states comply with international agreements under anarchy? These models emphasized the role of repetition, reputation, and hegemony in PD situations. However, these scholars were less successful in explaining how agreements are created.

SECOND-GENERATION MODELS

One reason why the neorealist–neoliberal debate was unproductive was that neorealists simply *assumed* that states maximized relative gains and then jammed those new payoffs into the four cells of the 2×2 PD. Second-generation scholars built models from the ground up that included the features of their intuitive arguments. They also modeled bargaining over international agreements and introduced domestic political actors.

The Vanishing Neorealist–Neoliberal Debate

Powell (1991, 1993) presented two models with important implications for the neorealist–neoliberal debate.[1] The first (Powell 1991) was a simple two-country model in which states decided whether to cooperate in the first period and then whether to attack in the second period. By assumption, states maximized absolute gains. However, if the costs of war were sufficiently low, then fear of attack in the second period could induce countries to worry about relative gains. This qualified the neorealist claims by showing that anarchy can induce a second-order concern about relative gains only if the costs of war are sufficiently low.

The second model (Powell 1993) was a two-state infinite-horizon model that examined the tradeoff between domestic consumption and military expenditures to defend against attacks. Although the model did not have explicit treaty-making, it addressed the question of whether anarchy is so dog-eat-dog that states must always forgo absolute for relative gains. Powell found many circumstances in which war did not occur and states could devote resources to consumption. He concluded that "the notion of anarchy has little if any significance distinctively related to international politics and . . . the problem of absolute versus relative gains is superfluous" (Powell 1993, p. 115). Given the dozens of articles on the neorealist–neoliberal debate, it is hard to overstate the importance of these conclusions.

Bargaining over International Agreements

First-generation models said little about the origins of cooperative agreements except that a hegemon can impose them on the global community. In first-generation models, states' utilities from cooperating are an exogenous parameter. In the real world, states' utilities from an agreement

[1]Niou & Ordeshook (1994) also offer a critique of the neorealist–neoliberal debate, but they do not present a model that addresses its shortcomings.

are endogenous outcomes from a bargaining game that occurs before states make compliance decisions. Second-generation scholars used bargaining models to derive endogenous compliance payoffs.

In Iida's (1993) model, two negotiators from the home and foreign state bargain over a one-dimensional issue space. Once agreement is achieved in the international arena, the negotiator from the home country submits the agreement to the home legislature. Ratification is decided by the median voter theorem—thus, the assumption of a one-dimensional bargaining space is crucial to the model. As Iida pointed out, this one-dimensional issue space is isomorphic to the contract curve of a larger-dimensional issue space, but a more interesting and informative model would have used the policy space itself rather than the contract curve as the bargaining space. Mo (1994) attempted to more explicitly model the bargaining between three different domestic constituencies. Two of the constituencies must vote for the agreement for it to be ratified. The foreign country makes the first proposal. If that proposal is rejected, then one of three domestic actors is chosen randomly to make a counter-proposal. Each proposer must craft proposals that are acceptable to two of the three domestic constituencies.

Fearon (1998) constructed a bargaining model (discussed in more detail below) that could easily be categorized as third generation, but we mention it here because of its implications for compliance. First-generation models show that discount rates must be sufficiently high for countries to comply with their agreements. Fearon used a war-of-attrition bargaining model to show that high discount rates also make bargaining more difficult, since states with higher discount rates will push harder for their most-preferred proposal. Patient countries are both more compliant and tougher negotiators. One limitation is that this result is not robust to the form of the bargaining model. The result does not necessarily hold in an incomplete-information, alternating-offers model.

Endogenizing compliance within a bargaining game raises an important selection problem. Presumably, states do not create international agreements that they know will generate non-compliance. We should only observe agreements that result in compliance. This confounds our ability to empirically observe the relationship between treaty design and compliance. A further implication is that states that have trouble complying with an agreement can use the threat of noncompliance to obtain a more favorable agreement in the bargaining stage. This point was recognized by Blaydes (2004), whom we discuss below.

The Complex Effects of Domestic Politics on Negotiations

Second-generation modelers also brought domestic politics into their models. Most were reacting to a paper by Putnam (1988) that used the term two-level games to describe the bargaining that occurs both at the international level, when states negotiate to create an agreement, and at the domestic level, when the executive negotiates with the legislature (or the electorate) to ratify the agreement. Putnam argues that domestic constraints can help a state's bargaining position in international negotiations. Having to placate a ratifying authority can give negotiators a better bargaining position. However, Iida (1993) and Mo (1994) use formal models to show that such claims cannot be generally supported.

Bringing domestic politics into models of international cooperation was an important theoretical innovation. However, the models became less useful for making empirical predictions. Whether domestic actors alter the international bargaining problem depends on specific model assumptions, including the information structure and the preferences of actors. Furthermore, the information that is available to actors in these models is almost always unobservable in practice, making it impossible to test these models. Although these models were useful heuristic exercises,

they were less useful for generating testable predictions about the role of domestic politics in international cooperation.

In addition to the articles above, Morrow (1991) used a model in which the executive can make arms control agreements to please the domestic electorate. The model predicts that arms control agreements would be more common in bad economic times because presidents would use them to show competence and increase their chances for reelection. However, if the economy is *very* bad, then the president will conclude fewer arms control agreements because doing so will not increase his chances of reelection (the economy is too bad for that) and will move defense policy further from his ideal point. Although the model did generate testable predictions, it was highly stylized and not at all general. This illustrates the strong assumptions that are necessary to generate clear, testable predictions in models with domestic politics.

THIRD-GENERATION MODELS

Major Theoretical Issues

First-generation models showed that it is possible for states to cooperate under anarchy by conditioning their behavior in each period on their opponent's behavior in prior periods. The key factor sustaining such cooperation is patience: states must care enough about the future that they are willing to resist the short-term temptation to defect in order to ensure long-term gains from cooperation. Second-generation modelers addressed the neorealist–neoliberal debate and introduced domestic politics and bargaining into their models. Third-generation scholars have begun to ask, "*How* should states cooperate?" This new wave of scholarship has been organized around four major questions:

1. How do states negotiate the distribution of cooperative gains?
2. What affects the depth of cooperation?
3. Why and how should agreements be designed to promote flexibility?
4. How does multilateralism affect the ability of states to cooperate?

We discuss each of these in turn.

Distribution. To understand the impact of distributional problems on international cooperation, consider the Organization of Petroleum Exporting Countries (OPEC). Cooperation in OPEC consists of restricting the supply of oil in the global market in order to increase its price. Each individual OPEC member has a short-term temptation to cheat by increasing its oil production unilaterally. Nevertheless, all OPEC members are better off if they jointly restrict supply rather than engaging in market competition. As Blaydes (2004) shows, OPEC faces a distributional problem. In addition to deciding how much aggregate oil to extract, the organization must also decide how much oil each member is allowed to extract. Although all states have an interest in keeping aggregate production low, each state prefers to receive a larger share of that aggregate production. This example shows that opportunities for cooperation can generate conflicts about how to distribute the benefits of cooperation.

Blaydes (2004) shows that patience has an important impact on the distribution of cooperative benefits. Her model has two stages. In the first stage, two states negotiate how to divide the benefits of cooperation. Blaydes assumes that a strong state makes a take-it-or-leave-it offer to a weak state. In the second stage, the two states play an infinitely repeated PD game. If both states cooperate, then the benefits of cooperation are divided according to the first-stage agreement. If an agreement is not reached in the first stage, then both states defect in all subsequent time periods. So the strong player never has incentive to make an offer that it knows will be rejected Recall that

patience is key to sustaining cooperation over time. If a state places low value on the future (i.e., is not patient), then it faces a strong temptation to defect in order to receive a short-term gain. As a player grows less patient, the payoff for joint cooperation must be increased in order for the state to be willing to cooperate. This means that the less patient the weak state is, the more the strong state must offer it in the first stage in order to achieve cooperation. Simply put, "impatience can help an actor to achieve a better outcome" during negotiations (Blaydes 2004, pp. 217–18).

Distributional problems are exacerbated when there is uncertainty about the preferences of players. Morrow (1994) adapts a one-period battle-of-the-sexes (BoS) framework to illustrate this point. In a BoS game, the worst outcome is a failure to coordinate actions. Both players are better off if they coordinate their choices. However, the two players differ in their most-preferred policy. This difference in preferences generates a mixed-strategy equilibrium in which players sometimes fail to coordinate. Morrow (1994) shows that allowing players to engage in cheap talk—i.e., send messages that are costless and nonbinding—can increase the likelihood that players will "cooperate" by choosing a common policy.

However, informational problems can have a more dramatic influence on distributional problems if strategic interactions are infinitely repeated. Fearon (1998) constructs a model in which two players must write an agreement that specifies how to divide up the gains from cooperation. Each state wants to receive a larger share. Then states play an infinitely repeated PD game. If both players cooperate, then the benefits of cooperation are divided according to the initial agreement. Fearon models the negotiation stage as a war of attrition. Such models are basically "staring contests" in which there is a costly fight and each player chooses a time at which to quit. A player loses if he quits the war ("blinks") sooner than his opponent. Each player has private information about his own instantaneous cost of delay. Fearon shows that the player with the lower cost of delay is always willing to wait longer than his opponent and hence receives his most-preferred outcome. The key result of the paper pertains to the role of patience. Fearon finds that as "states care more about future payoffs . . . , all types choose tougher bargaining strategies" (1998, p. 283). That is, all states—regardless of their own individual cost of bargaining—will bargain harder as they grow more patient. Although patience is crucial to the ability of states to sustain cooperative agreements, it also exacerbates distributional problems. When states care more about the future, they are more willing to pay short-term costs to secure agreements that give them better terms. This does not necessarily reduce the efficiency of cooperation because a larger discount factor means that future cooperative benefits are more valued by the players. Nonetheless, more patience makes it more difficult and more costly to reach an initial cooperative agreement.

Depth. An enduring debate among IR scholars is how much international organizations (IOs) actually change behavior. Many scholars focus almost exclusively on compliance: the extent to which "actual behavior of a given subject conforms to prescribed behavior" (Young 1979, p. 140). Optimists argue that "almost all nations observe almost all principles of international law and almost all of their obligations almost all the time" (Henkin 1968, p. 47). Thus, they conclude, international law and organizations are effective in promoting cooperation. However, rational choice theorists have emphasized that a treaty or institution is effective only if it changes state behavior. For example, suppose that, absent a trade agreement, the United States will find it optimal to set a tariff rate of 20% on cotton. Suppose a trade agreement specifies that the United States must keep its cotton tariff below 25%. Two things are obvious. First, the United States will fully comply with this agreement. Second, the existence of the agreement will not change the behavior of the United States, since its most-preferred tariff of 20% is permissible under the treaty. Compliance does not necessarily imply effectiveness.

Downs et al. (1996) emphasize this point using a simple model of international trade. The key variable in their model is the depth of the international agreement: "the extent to which [an agreement] requires states to depart from what they would have done in its absence" (Downs et al. 1996, p. 383). They show that as a trade agreement requires deeper tariff reductions, states have a greater temptation to violate it. If the treaty creates only slight constraints on state behavior, then a country will be very likely to comply. However, this treaty will not be very effective because it requires little of its signatories. On the other hand, if the treaty imposes very strong constraints on state behavior, then a country is less likely to comply. Downs et al. argue that the level of enforcement is a critical factor in the effectiveness of international cooperation. Deeper treaties will require more enforcement in order to induce compliance.

This tension between depth and compliance is also highlighted by Johns (2011b). In this model, two trading states face stochastic shocks over time in their need to protect domestic import-competing industries. A trade agreement specifies both the depth of cooperation—the maximum permissible tariff rate—and the size of the fine that a state must pay its trading partner if it violates the treaty but wishes to remain a member of the treaty in the future. The payment of this fine is voluntary. Larger fines are essentially higher levels of enforcement. The key result of the model is that when states design a treaty, they face a tradeoff between the depth of the treaty and the optimal level of enforcement. Compliance concerns alone might suggest that states should always design strong enforcement systems to support deep treaties. However, there is often good reason for states to design agreements that do not always yield high levels of compliance and effectiveness. This concern is most apparent in the literature on flexibility in international agreements.

Flexibility. As discussed above, early models of IOs frequently assumed grim trigger punishments, in which a single defection results in a loss of cooperative benefits for all future periods. This can ensure cooperation if players are patient enough to sacrifice the short-term gain from defection in return for the long-term benefits of mutual cooperation. However, the dynamics of cooperation change if players are uncertain about the costs and benefits of cooperation.

One strand of the flexibility literature argues that the pressures that leaders face to defect on cooperative agreements can vary stochastically over time. For example, a leader who is facing an imminent election or bad economic conditions may be under pressure to protect domestic industries by violating trade commitments. One way that agreements can be designed to account for such political pressure is to build in escape clauses. As defined by Rosendorff & Milner (2001, p. 830), "[a]n escape clause is any provision of an international agreement that allows a country to suspend the [commitments] it previously negotiated without violating or abrogating the terms of the agreement." In all of these models, there is a cost to violating the agreement. In order to remain a member of the cooperative regime, a defecting state must pay a voluntary fine. The size of this fine is a measure of the flexibility of the agreement. Very flexible agreements assign relatively small fines to violations, whereas very rigid agreements assign large fines.

Downs & Rocke (1995, ch. 4), Rosendorff & Milner (2001), and Rosendorff (2005) all find that the inclusion of escape clauses enhances the stability of cooperative agreements. Under a grim trigger punishment, if a leader faces intense political pressure and violates an agreement, then the cooperative regime collapses. In contrast, if an agreement contains an escape clause, which permits tolerated defection under certain circumstances, then the cooperative regime is more likely to remain in place over time. These authors also argue that escape clauses make it more likely that an agreement will be formed. As shown by Blaydes (2004) and Fearon (1998), the temptation of states to cheat on an agreement affects the initial negotiations between states about the form of the agreement. According to Rosendorff & Milner (2001) and Rosendorff (2005), escape clauses make it easier for states to negotiate agreements because states will be more likely

to accept an agreement if they know that they can violate it during tough times. Johns (2011b) builds on this framework to show that more flexible agreements—which assign relatively small fines to defections—are more likely to specify deeper levels of cooperation. For example, if a leader knows that he has the flexibility to violate a trade agreement when times are tough, then he will commit to deeper initial tariff concessions. The overall effect of flexibility on average levels of cooperation is ambiguous.

A second strand of the flexibility literature studies the impact of uncertainty about the size of the benefits that a particular cooperative agreement will yield for each state. Rather than focusing on domestic political or economic pressure that is external to the treaty, this literature focuses on uncertainty about the treaty itself. Koremenos (2001, 2005) argues that an agreement that is initially acceptable to two states may no longer be acceptable if one state unexpectedly gains a small share of the benefits or a large share of the costs. For example, an agreement to lower carbon emissions can be destabilized if a treaty member unexpectedly faces very high costs of compliance. One way to ameliorate such instability is to design agreements of limited duration or to explicitly allow for renegotiation. This increases contracting costs because states can or must return to the bargaining table multiple times to adjust the allocation of costs and benefits. However, limiting the duration of an agreement or allowing for renegotiation introduces flexibility into the design of the treaty, which can facilitate compliance with treaty terms. Additionally, this flexibility makes risk-averse states more willing to join a regime in the first place.

Multilateralism. Another area in which formal models are increasingly being used is in understanding multilateralism. There is usually a presumption that larger IOs face special challenges in promoting international cooperation because member preferences are more diverse. For example, members of the North Atlantic Treaty Organization (NATO) have more homogenous preferences on national security matters than do members of the United Nations (UN). Multilateral cooperation has many possible benefits over bilateral cooperation, including the opportunity for a greater impact on an issue, sharing of technical expertise, and the increasing returns to scale that come from centralized administration (Abbott & Snidal 1998). However, multilateral cooperation also comes with added costs. As emphasized by Milgrom et al. (1990), monitoring cooperative behavior within a large community of individuals is much more challenging than simply monitoring bilateral interactions. States face a greater temptation to cheat and free ride on large multilateral agreements if they are less likely to be caught.

Another possible cost is that more diversity in state preferences might increase the difficulty of negotiating an initial agreement about how cooperation should take place. This is commonly known as the broader–deeper tradeoff: organizations with a broader set of member preferences may be less likely to reach deep agreements than organizations with a narrower set of member preferences. For example, NATO may be more effective in promoting security cooperation than the UN because the membership of NATO is smaller and more homogenous.

Downs et al. (1998) argue that one way the broader–deeper tradeoff can be ameliorated is through the gradual growth of an institution over time. In their account, all members of an IO must commit to a common policy that is chosen through internal voting. Suppose a group of states wishes to promote European integration, but the states differ in their preferred level of integration. If all states join an IO at the same time, then Downs et al. (1998) find that the chosen policy will lie at or near the ideal policy of the median IO member. Such an inclusive organization will choose a moderate level of integration.

In contrast, suppose a small organization is formed of only those states that favor very deep integration. The policy choice that arises from this group will be much more pro-integration than the policy chosen by the inclusive organization. If this small group adopts a supermajoritarian

decision rule—such as requiring a two-thirds vote to overturn a status quo policy—then the slow addition of new, more conservative members over time will have only a modest impact on making the IO policy less integrationist. This is because the supermajoritarian decision rule biases outcomes toward the initial status quo policy, which was chosen by pro-integration states. Downs et al. refer to this as "managing the evolution" of the IO because "all else equal, large multilaterals that start out small will be able to achieve considerably more depth than those that start out relatively large" (1998, pp. 397, 414). As we discuss below, though, this result depends crucially on the assumption that states set their policies at the same level. In a related model, Schneider & Urpelainen (2011) examine the EU's unanimous accession rule, which requires all current members to agree in order for a new member to be admitted. They argue that prospective members will make more policy concessions if they are required to gain the support of the full membership rather than a simple majority. These models focus on promoting cooperation. However, "cooperation" is not necessarily an unmitigated good. After all, cooperation within OPEC entails raising oil prices, which hurts consumers and nonmember governments. One of the few papers to model such negative externalities is by Kydd (2001). In this model, NATO members decide whether to expand to include Eastern European states. This might facilitate cooperation among member states. However, one cost is that Russia, which is uncertain about the intentions of NATO and Eastern European states, may interpret expansion as an aggressive act. This model shows that members of an existing multilateral IO can sometimes use costly screening procedures to both (*a*) ensure that new members are supportive of the institution's goals and (*b*) signal benign intentions to nonmembers who fear negative externalities. Imposing conditionality and accession costs on new members can be essential in building effective institutions.

Verdier (2008) examines the combination of multilateralism and bilateralism. This model of nuclear proliferation agreements shows that multilateral regimes can actually be complements, rather than alternatives, to bilateral regimes. Verdier assumes that a "regime-designer" is trying to get a group of states to adopt a particular behavior. A key factor is contracting costs. Multilateral agreements are relatively cheap to write because they specify a single set of rules about expected behavior (e.g., "don't build nuclear weapons"), rewards, and sanctions. However, many multilateral members receive far more benefits from the regime than are necessary for them to be a member. For example, many states would likely never develop nuclear weapons even if the Non-Proliferation of Nuclear Weapons Treaty (NPT) did not exist. The existence of the NPT means that these states can gain the benefits of treaty membership (such as assistance in the development of peaceful nuclear technology) without making any sacrifices. In contrast, bilateral agreements can be customized so that each member receives just enough benefits to be willing to cooperate, and no more. However, it is costly to write these individual agreements. Rather than choosing between a bilateral and a multilateral regime, Verdier studies a hybrid regime of both multilateral and bilateral agreements. States that find compliance very cheap are offered minimal rewards for compliance. States that find compliance more difficult are offered bilateral agreements with more generous rewards. The regime-designer prefers this hybrid system to a purely multilateral or bilateral system of contracts.

A core assumption driving most accounts of multilateralism is that all members must choose a common policy. However, Gilligan (2004) shows that IOs can be both broad and deep if they allow members to adopt different policies. This is particularly apparent in multilateral trade agreements. The General Agreement on Tariffs and Trade/World Trade Organization (GATT/WTO) requires members to abide by certain common legal principles, such as granting other members most-favored nation status. However, the GATT/WTO system also allows its members discretion over specific tariff obligations, which are negotiated on a product-by-product basis. Each state can balance its concessions across a broad range of products in response to its domestic political needs.

A member that finds it politically prohibitive to liberalize on cotton, for instance, can instead offer deeper concessions on another product, like steel. Additionally, the GATT/WTO has different rules and exceptions for less developed countries. Not all states can liberalize at the same level on every product. By allowing members the ability to choose different policies, the GATT/WTO is both broad and deep.

Functions of International Organizations

Third-generation formal models have also examined the functions of IOs. This literature can be loosely organized around four uses of IOs: authorizing the use of force, manipulating domestic politics developing bureaucratic expertise, and adjudicating disputes. We discuss each of these in turn.

Authorizing the use of force. Most formal models of the UN focus on the Security Council's ability to authorize the use of force by UN members. Although the UNSC lacks enforcement capabilities, many qualitative accounts argue that its decisions can signal important information to domestic constituencies about whether a coercer is proposing a good policy (Thompson 2006, Voeten 2005). Recent formal models support these arguments. These models focus on principal–agent relationships between voters and political leaders over the conduct of foreign policy. All of these models begin with the assumption that leaders have more precise knowledge than voters about the likely consequences of foreign policy choices. Fang (2008) finds that state behavior at the UN is driven largely by pooling dynamics. In this paper, Fang assumes that the policy preferences of the international institution match those of the voter: force can sometimes be a good policy and sometimes not. "Good" leaders, i.e., those whose preferences match those of the voters, will seek UN authorization to try to convince voters that the use of force is necessary. However, "bad" leaders, who prefer force regardless of whether it is necessary, mimic the behavior of the good leaders by also seeking UN authorization. The primary value of the institution is to signal information about the quality of policy choices, which in turn allows voters to infer the quality of their leader.

Chapman (2007, 2011) makes a similar argument using a spatial model in which the leader and the pivotal member of the international institution have private information that is unavailable to the voter. This model allows for the possibility that the pivotal member of the institution may have an ideal policy that differs from that of the voter. Chapman concludes that seeking authorization from an IO can be most informative when the institution is biased in its preferences. Leaders have incentive to (*a*) seek approval from institutions that oppose their policy preferences and (*b*) avoid disapproval from institutions that support their policy preferences. The logic of Chapman's argument can be illustrated by comparing the UNSC and NATO. Generally speaking, NATO is more amenable to U.S. policies than the UNSC. If NATO—the "yes man"—supports the U.S. leader's request to use force, then voters are unlikely to change their beliefs significantly about whether force is a good policy. However, if NATO rejects the request, then voters can infer: "If the 'yes man' says 'no,' then using force must be a very bad decision." In contrast, decisions by the UNSC—the "no man"—signal the most information if they support the use of force. The voter reasons: "The UNSC really dislikes the use of force, so if they support the use of force in this case, then it must be necessary."

The incentives that are highlighted by Chapman (2007) bear a marked resemblance to the tradeoff between type I and type II errors in the design of medical diagnostics. The voter is uncertain about whether the use of force is a good decision. A conservative institution such as the UNSC will set a high bar for authorizing the use of force. This will minimize the likelihood

of false positives—saying that force is appropriate when it actually is not. But it will increase the probability of false negatives—not supporting the use of force when it is in fact beneficial. In contrast, a more activist institution such as NATO is less likely to yield a false negative but more likely to yield a false positive. It is not possible to design a single decision rule that minimizes both false positives and false negatives (Greene 2008, p. 1035; Hogg et al. 2005, p. 265). However, the existence of many security institutions allows for diversity in the standards used to authorize the use of force. We suspect that the ability of leaders to forum-shop across these different institutions leads to better outcomes with regard to both the policies chosen by a leader and the ability of voters to infer the true preference of their leader. This effect might help to ameliorate some of the concerns of critics of forum-shopping and regime complexity (Drezner 2009, Raustiala & Victor 2004).

The overall effect of IO authorization in this class of models is to lower the costs of coercion by increasing domestic and/or international support for the use of force by a challenger state. Chapman & Wolford (2010) show that this dynamic affects crisis bargaining. In their model, a challenger state makes a take-it-or-leave it offer to the target. If the target refuses the offer, conflict ensues. Chapman & Wolford assume that players are uncertain about the resolve of the target state and that the IO is biased either for or against the challenger for exogenous reasons. So when the challenger is deciding whether to seek IO authorization, he does not need to worry about whether this choice signals information about his competence to voters. If IO authorization lowers the challenger's cost of conflict, then the challenger will have incentive to make larger demands, which the target will be less likely to accept. So IO authorization increases the probability that a war will occur. IO opposition to the use of force has the converse effect: it raises the challenger's cost of war, which leads to smaller demands that are more likely to be accepted prior to conflict. However, if the challenger anticipates that the IO will oppose its request to authorize the use of force, then it is unlikely to go to the IO in the first place. By the logic of Chapman & Wolford (2010), an IO is least likely to be used when it is most likely to be effective in promoting peaceful settlement of disputes. Although IOs may increase transparency about the quality of policy decisions to voters, they can also have the perverse effect of increasing the likelihood of war.

Manipulating domestic politics. The above analyses all relied on an informational mechanism: a leader sought IO approval to signal information to voters about the quality of his foreign policy decision. This signaling mechanism is not unique to security IOs. Mansfield et al. (2002) argue that leaders use international trade agreements to signal economic competence to domestic constituencies. They assume that a leader and his voters will always have conflicting preferences over trade policies. The leader prefers high tariffs, which garner him rents from protected industries, whereas voters prefer low tariffs, which lower consumer prices. In their baseline model without a trade agreement, voters know the prices of products but do not directly observe tariff policies. This means that if consumer prices are particularly high, voters will be uncertain about whether this was caused by high tariffs or by adverse economic shocks that are beyond the control of the leader. Voters must decide whether to support or oppose the leader. However, sometimes a "mistake" will be made: a leader who has chosen a low tariff will be opposed because of a bad shock. According to Mansfield et al., dispute settlement mechanisms—procedures by which one trading partner can accuse another of violating a treaty obligation—are key to increasing transparency about government policies. They argue that voters "are more likely to hear about a foreign government's or international organization's complaints regarding their government's violations of a trade agreement than they are to learn about changes in domestic trade policy" (Mansfield et al. 2002, p. 480). This means that the existence of a trade agreement provides voters with information about the policy chosen by their leader. The overall effect of the institution is to lower tariff rates.

Nonetheless, the leader benefits from this enhanced transparency because it lowers the likelihood that he is punished for bad economic shocks that are beyond his control. If the country is sufficiently democratic, then the benefits a leader receives from transparency outweigh the costs.

However, IO involvement does not always imply that a leader has benign intentions or is choosing policies that are good for his domestic constituents. In a recent provocative paper, Hollyer & Rosendorff (2011) argue that membership in human rights treaties can actually increase a brutal autocrat's hold on political power. In this model, a domestic opposition group is contemplating whether to take a costly action against an autocrat, such as organizing protests, rallies, and other public challenges. The opposition group is uncertain about how much the leader values remaining in office. Leaders who place a high value on remaining in office will be more likely to respond violently to opposition activities than leaders who place less value on remaining in office. Hollyer & Rosendorff assume that membership in a human rights treaty increases a leader's cost of suppressing his political opposition. This creates perverse incentives because the leader's willingness to bear these increased costs can be interpreted as indicating that he is very resolved to stay in power. They argue that "authoritarian states sign human rights treaties explicitly because they do *not* intend to comply. . . . The signing of a human rights treaty is a signal to the opposition of the high value the elite places on holding onto power and its willingness to use torture" (Hollyer & Rosendorff 2011, p. 3). When the political opposition sees that the leader has signed the treaty, it decreases its level of effort and the leader is more likely to survive in office.

Scholars often describe IOs and agreements as commitment devices that "tie the hands" of leaders. By increasing the cost of particular actions—be it raising tariff levels or expropriating foreign investment—they presumably make a leader less likely to engage in these actions. However, Hollyer & Rosendorff (2011) show that sometimes constraints can be interpreted in the opposite manner. If two men are arguing, the willingness of one to raise the cost of violence might be interpreted as signaling that he does not intend to fight. However, it can also be interpreted as signaling that he is so strong that he is willing to fight with one arm tied behind his back. This reinforces a point made earlier by Nalebuff (1991, p. 316): "[t]he value of a reputation depends on how others interpret it." Although actors may interpret a costly action—be it signing a human rights treaty or seeking UNSC approval—as signaling good intentions, it is often possible to construct equilibria in which such actions signal bad intentions.

One final strand of the literature on IOs and domestic politics emphasizes the impact of international politics on the creation or empowerment of domestic constituencies that favor cooperation. For example, most political economy accounts of trade policy emphasize the importance of lobbying by import-competing industries against trade liberalization (e.g., Grossman & Helpman 1994). Economic theory suggests that the gain to consumers from free trade outweighs the costs to import-competing industries. However, the benefits of free trade are diffused over all consumers, while the costs are concentrated. The logic of collective action problems suggests that import-competing industries will be more active and successful in lobbying for trade protection than consumers will be in lobbying against it. This results in trade policies that are suboptimal for the economy as a whole. Gilligan (1997) argues that one important impact of the modern system of reciprocal trade negotiations—in fora such as the GATT/WTO—is the empowerment of exporting industries. These actors care most about the tariffs chosen by other countries. They have no direct incentive to lobby their home government for lower tariffs in their own market. However, since the modern trade system relies on reciprocal negotiations and commitments, exporters have incentive to lobby for trade liberalization at home because they know this will translate to lower tariffs in foreign markets. The structure of modern trade negotiations enhances the incentives for procooperation constituencies to lobby their government.

Dai (2005) develops a similar theoretical argument in the context of environmental cooperation. She argues that the willingness of political leaders to join and comply with an environmental treaty is affected by the lobbying activities of domestic special-interest groups. While some groups (such as manufacturers) are likely to oppose cooperation, other groups (such as civil society organizations) are likely to favor it. The ability of these groups to influence policy depends on two factors: electoral leverage and issue expertise. Dai notes that one way IOs can promote cooperation is "by increasing the electoral leverage and improving the informational status of procompliance constituencies. In doing so, international institutions can facilitate a decentralized compliance system, where the enforcement source is not from some states over others, but rather from some domestic constituencies over their government" (2005, p. 366). For example, commissioning and publicizing studies about an environmental problem can build the electoral leverage of procooperation groups by convincing voters that politicians should protect the environment. Similarly, information sharing at the international level can help domestic interest groups to learn the likely effects of various policy actions and to monitor their government's behavior.

Dai's work points to one way international institutions can incentivize domestic actors to punish leaders who cheat on agreements. McGillivray & Smith (2000, 2004, 2006, 2008) offer a different model with a similar outcome. First-generation scholars used trigger strategies to explain why states comply with their international agreements. Two decades later, McGillivray & Smith gained new and important insights by introducing models with "leader-specific punishments." In their models, punishments target the leader who cheated on the agreement. These punishments are removed once the cheating leader is removed from office. McGillivray & Smith show that this enhances the enforcement power of trigger strategies.

Developing bureaucratic expertise. Another key function of IOs is the development of bureaucratic expertise. This is most apparent in relatively technocratic institutions, such as the International Monetary Fund (IMF) and the WTO. However, even the UN—an inherently political body—is associated with a vast bureaucracy that is responsible for tasks such as promoting economic development, administering public health programs, and inspecting nuclear weapons programs. Some recent publications use cheap talk models to examine the relationship between politicians and international bureaucrats.

In cheap talk models, the bureaucrat (the "agent") has knowledge about how policy choices translate into final outcomes. The bureaucrat can send a costless message to a politician (the "principal"), who chooses policy. However, this bureaucrat does not have incentive to communicate truthfully if he believes that the politician's preferences differ from his own. Scholars in the American politics subfield have used this framework to study communication between policy experts and legislators. They find that messages sent by the expert are most informative when the expert's policy preferences closely match those of the median legislator (Gilligan & Krehbiel 1989). This suggests that if a legislature wants a bureaucratic agency to fully share its expertise, then the agency must be staffed by individuals whose preferences are closely aligned with those of the median legislator.

However, IOs rarely look like domestic legislatures for two reasons. First, the median voter theorem does not apply in most IOs. Second, decisions within IOs can rarely be considered binding, since member states can always act unilaterally outside of the institution. Johns (2007) examines a cheap talk model that accounts for both of these attributes of IOs. The model assumes that the bureaucrat reports to two principals, who must bargain over which policy to choose. After the report is made, each principal is able to unilaterally leave the bargaining table and exercise an outside option. The key result is that biased bureaucrats can sometimes make all IO member states better off than a moderate bureaucrat. If one state can credibly threaten to leave the bargaining

table and take unilateral actions, then policy choices will more closely reflect the preferences of this state (Voeten 2001) rather than having "moderate" outcomes. A bureaucrat will be most willing to reveal his private information if he likes these biased policies. All politicians in this model benefit from more precise information about how policies translate into outcomes. Johns (2007) provides a theoretical justification for why international bureaucracies should be designed to reflect the preferences of states that can credibly threaten unilateral action.

In a related piece, Fang & Stone (2011) examine the interaction between international and domestic bureaucrats. A domestic leader may be uncertain about what kinds of economic or public health policies are appropriate for her country. This leader has two choices. First, should she seek the advice of an international bureaucracy such as the IMF or the World Health Organization? Second, should she delegate her decision-making authority to a domestic policy expert, who has his own information about what policy is appropriate? The IO is a political actor whose preferences can diverge from those of domestic policy makers. The IO's willingness to share information will depend on how this information influences policy choices. Fang & Stone find that an international bureaucracy has an impact on policy outcomes only if the IO is a moderating force between the politician and the domestic policy expert.

Finally, Urpelainen (2011) argues that the international bureaucrats are susceptible to countervailing pressure by IO member states. If a bureaucrat can choose and implement policies independent of member-state approval, then IO members have incentive to use resources to influence his decision. Urpelainen holds that states may deliberately limit the influence of international bureaucrats because the costs of competing for influence outweigh the benefits of developing bureaucratic expertise.

Adjudicating disputes. IOs can also play an important role in adjudicating disputes. Recent decades have seen a dramatic proliferation of courts, such as the International Court of Justice, the Dispute Settlement Mechanism of the WTO, and the International Criminal Court (ICC) (Romano 1998). Legal scholars have developed complex taxonomies to classify and describe international judicial bodies (e.g., Romano 2011). For the sake of parsimony, we refer to these institutions generally as "courts." Studies of domestic courts usually focus on courts as "deciders" of contentious issues. These accounts assume that court rulings are enforced and the court has jurisdiction to rule. IOs appear quite weak from this perspective because compliance with court rulings and membership in international courts is voluntary. Even courts that view their ruling as binding must rely on the willingness of states to comply (Bello 1996). Nevertheless, recent formal models have emphasized ways in which international courts can be effective in resolving disputes.

Many models focus on courts as information providers. As described above, Mansfield et al. (2002) argue that the dispute-settlement procedures in trade agreements create domestic-level transparency about trade policy. This can be beneficial to both voters and leaders if a country is sufficiently democratic. Carrubba (2005, 2009) argues that international courts can observe or infer a state's cost of compliance with an international regulatory regime. Recall that many flexibility theorists argue that it is important for IOs to tolerate defection when the costs of compliance are excessive. Carrubba shows that if judges are strategic actors who seek to promote compliance, then they will be willing to rule against a defecting state only if they believe that the state will comply with the ruling by changing its behavior. Accepting judicial review is analogous to joining a flexible regime in which states can defect from their treaty obligations during tough times.

However, Johns (2011a) demonstrates that courts can be effective in resolving disputes even if they provide no informational value to disputants. In this model, two states are involved in a

dispute and the court randomly determines who wins the case. The court's ruling is nonbinding and disputants can continue to negotiate after a ruling is made. Johns shows that even if the court is nothing more than a "coin-flipper," the court's ruling can coordinate costly enforcement by third parties, who have no inherent interest in the dispute. This transforms a bilateral dispute into a multilateral problem. These third-party states have incentive to provide costly enforcement if they believe they will be involved in future disputes that can be referred to the court. This highly abstract model shows that international courts do more than just decide issues or provide information to disputants. They also coordinate multilateral involvement in bilateral disputes. Fang (2010) argues that the prospect of such enforcement can influence bargaining interactions between states even if the court is not actually used, and Reinhardt (2000) contends that uncertainty about such enforcement affects the size of pretrial concessions.

One additional way in which courts can influence international politics is by solving commitment problems. Gilligan (2006) examines the role of the ICC. He argues that, over time, despotic leaders experience shocks in their ability to maintain power. Repression can be useful in suppressing dissent. However, sometimes a despot is tempted to step down from power if he knows that he can receive asylum in a foreign country rather than being killed by rebels. Prior to humanitarian violations, a foreign state that supports human rights is tempted to tell the despot, "If you repress your people, then I will never give you asylum." If credible, this threat lowers the expected utility of repression because the despot knows that he may want to flee his country when times get tough. However, if atrocities actually are committed, then the foreign state would rather the despot step down from power than continue to repress his people. Absent a court, the foreign state's threat is not credible because the state cannot commit to refusing asylum to leaders who abuse human rights. The ICC issues arrest warrants, but it has no police force to arrest a political leader. Gilligan argues that when a despot is highly likely to be overthrown, he would rather surrender to the ICC than face death or prosecution by his domestic political opponents. In such situations, the foreign country can refuse an asylum request because it knows that the leader is willing to step down from power and face international prosecution. The ICC solves the foreign state's commitment problem because it allows third parties to credibly commit to refuse asylum to repressive leaders. Ritter & Wolford (2011) argue that if the ICC were to engage in prearrest bargaining, then it could propose smaller punishments for those leaders who are less likely to be removed from power. This lenience would increase the ability of the ICC to prosecute crimes because despots would be more apt to surrender. However, it would also increase the likelihood of humanitarian crimes, since prearrest bargaining effectively lowers the expected punishment for crimes. This suggests a tradeoff between the ICC's objectives of punishing and deterring.

All these models discuss ways in which international courts solve strategic problems. However, recent scholarship argues that stronger courts—with higher levels of enforcement and jurisdiction—are not necessarily an unmitigated good. The legalization of international politics comes with hidden costs. Both Gilligan et al. (2010) and Johns (2011c) contend that when two states are involved in a dispute, the plaintiff has private information about the likelihood that he will prevail in litigation. If the court is relatively weak—because either its rulings have little impact on final outcomes or it is unlikely to assert jurisdiction over the case—then the legal claims are not very important. After all, if the law doesn't matter, then there is little reason for two states to fight over differences in legal opinions. However, as the court grows stronger, legal claims matter more, and the asymmetric information becomes more important. This creates bargaining problems between the two states. The overall impact of these problems is to lessen the likelihood of an early settlement to the dispute and increase the likelihood of costly litigation. Strong courts can have the perverse effect of exacerbating conflicts between states.

CONCLUSION

To distill 30 years of scholarship down to a few simple points is a daunting task. Instead of trying to be exhaustive in this conclusion, we summarize the main lessons that we think formal models of international institutions have taught us. We divide these insights into lessons taught by first-, second-, and third-generation models. For the latter, we identify a trend toward institution-specific research and highlight some questions that we hope will guide the next generation in its work.

First-generation models of international cooperation are concerned mainly with the question of why states comply with their cooperative agreements under anarchy. These models rely on the iterated PD. As long as discount rates are sufficiently high, trigger strategies can make agreements self-enforcing. Institutions can help states to comply with their agreements even when those institutions are seemingly toothless. Medieval law merchants identified cheating traders so that those traders would face multilateral sanctions for their transgressions. International courts and dispute resolution mechanisms play a similar role today. These models have shown that individual actors can have incentive to punish cheaters even when they were not the party that was cheated.

Neorealists criticized these new theories by arguing that they missed the most important feature of anarchy: states are responsible for their own security. This implies that cooperation, if not outright impossible, is much harder than cooperative models suggest. Although the resulting neorealist–neoliberal debate consumed more journal pages than was perhaps necessary, it did lead to an important truth: states with a monopoly on the use of force are not necessary for efficiency, order, or even security. Formal models have released IR scholarship from the belief that the threat of war is always present and therefore cooperation too risky. Instead, these models have shown that states in the international system have no incentive to engage in violent conflict when the cost of such conflict, broadly understood, exceeds its expected benefits. When this is true, war is effectively taken off the table, thereby opening up opportunities for mutually beneficial cooperation. It is possible for states to negotiate and comply with efficiency-enhancing cooperative agreements even in a state of anarchy. Various second-generation bargaining models examine how states create these agreements, and the repeated PD and reputational concerns tell us why states comply with them.

This second-generation literature also produced some valuable lessons about the role of domestic politics in the creation of international agreements and the effects of international institutions on the distribution of domestic political power. Almost 60 years ago, Waltz (1954) warned that using domestic politics to explain international politics came at a terrible price in terms of parsimony. To some extent, international bargaining models that include domestic politics have corroborated that concern. As simple and stylized as these models are, they are still unable to offer a simple answer to the question of whether domestic constraints help or hurt an international negotiator. This indeterminacy is not the fault of the modelers or their models but simply a fact of modeling: when more actors and incomplete information are added, results become increasingly case-contingent.

Third-generation models that study the effect of international institutions on domestic political power have had a bit more luck. They show that international cooperation can sometimes create its own advocates by concentrating the benefits of cooperation and thereby alleviating collective action problems among domestic supporters of cooperation.

Our sense is that the contemporary literature on international institutions is growing increasingly institution specific. Rather than focusing on broad theoretical issues—multilateralism, distribution, depth, etc.—most contemporary research focuses on particular issue areas and institutions. Even the flexibility literature, which has broad implications for the design of international cooperation, is currently grounded in international trade applications. We think that this trend comes

with both benefits and costs. A benefit is that specialization may facilitate communication and collaboration with empirical scholars, who tend to focus on specific issue areas (trade, investment, etc.). However, this trend also opens the door to fragmentation within the literature. It is very easy, for example, for a young theorist to focus exclusively on the literature related to his or her issue area and ignore developments in other areas. Yet formal models play an important role in elucidating broader mechanisms that drive behavior. There seems to be a broader–deeper tradeoff in contemporary research. The push toward deeper models with testable empirical implications and institution-specific explanations means that we often lose sight of the broader implications of contemporary work. This presents a challenge for future scholars.

Earlier generations showed the need for cooperation and the ability to create governance from anarchy. However, most past conceptions of international institutions have been relatively monolithic. Recent research has suggested the importance of treating these institutions—and the people who work within them—as strategic actors. There is no reason to assume that international institutions and their employees are impartial actors. A few third-generation papers examine the impact of an international bureaucracy that pursues its own interests instead of those of the states that created it. One of the big challenges for such researchers is to specify the preferences of international bureaucrats. Are they policy-motivated or career-motivated? If they are career-motivated, how does one model such actors? Once these hurdles are overcome, formal modelers will be better equipped to study the effects of self-interested international institutional actors. This line of research has potentially important normative implications.

The recent formal work that we find most provocative and innovative highlights the negative consequences of international institutions. It is almost an article of faith among scholars of international institutions that cooperation is an unmitigated good. However, recent models show that international institutions can also generate perverse outcomes. For example, membership in human rights agreements can reduce political opposition to nasty leaders because the very act of increasing the costs of repression can signal a high willingness to repress. Similarly, strong international courts can hinder the settlement of international disputes by increasing the importance of uncertainty about legal claims. Finally, while cooperation may benefit the members of a privileged group—be it a military alliance, a trade agreement, or a commodity cartel—it can harm nonmembers. It is worth remembering that "cooperation" in a PD consists of prisoners avoiding punishment for their crimes, which is surely not beneficial to society. We think future work that analyzes the possibly pernicious side effects of cooperation and institutionalization is particularly important for public policy. If international cooperation is in some cases one more way for the strong and well-connected to exploit the weak and disadvantaged, then we would hope that formal modelers of international cooperation help identify those cases. Three generations of scholarship have used formal models to explore the ability of states to create order from anarchy. We hope that future scholars will continue this work while being cognizant that order is not always a good thing.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank James Fearon for proposing the idea for this article and for helpful suggestions on its organization and content. We also owe a debt of thanks to Lauren Peritz, who provided excellent research assistance, and Peter Rosendorff, who provided feedback on an early draft. All errors remain our responsibility.

LITERATURE CITED

Abbott KW, Snidal D. 1998. Why states act through formal international organizations. *J. Confl. Resolut.* 42:3–32

Alt JE, Calvert RL, Humes BD. 1988. Reputation and hegemonic stability: a game-theoretic analysis. *Am. Polit. Sci. Rev.* 82:445–66

Axelrod R. 1985. *The Evolution of Cooperation*. New York: Basic Books

Bello JH. 1996. The WTO dispute settlment understanding: less is more. *Am. J. Int. Law* 90:416–18

Bendor J, Swistak P. 1997. The evolutionary stability of cooperation. *Am. Polit. Sci. Rev.* 91:290–307

Binmore K. 1998. Review of R. Axelrod, *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. *J. Artif. Soc. Soc. Simulation*

Blaydes L. 2004. Rewarding impatience: a bargaining and enforcement model of OPEC. *Int. Organ.* 58:213–37

Carrubba C. 2005. Courts and compliance in international regulatory regimes. *J. Polit.* 67:669–89

Carrubba CJ. 2009. A model of the endogenous development of judicial institutions in federal and international systems. *J. Polit.* 71:55–69

Chapman TL. 2007. International security institutions, domestic politics, and institutional legitimacy. *J. Confl. Resolut.* 51:134–66

Chapman TL. 2011. *Securing Approval: Domestic Politics and Multilateral Authorization for War*. Chicago: Univ. Chicago Press

Chapman TL, Wolford S. 2010. International organizations, strategy, and crisis bargaining. *J. Polit.* 72:227–42

Coase R. 1960. The problem of social cost. *J. Law Econ.* 1:1–44

Dai X. 2005. Why comply? The domestic constituency mechanism. *Int. Organ.* 59:363–98

Downs GW, Rocke DM. 1995. *Optimal Imperfection? Domestic Uncertainty and Institutions in International Relations*. Princeton, NJ: Princeton Univ. Press

Downs GW, Rocke DM, Barsoom PN. 1996. Is the good news about compliance good news about cooperation? *Int. Organ.* 50:379–406

Downs GW, Rocke DM, Barsoom PN. 1998. Managing the evolution of multilateralism. *Int. Organ.* 52:397–419

Drezner DW. 2009. The power and peril of international regime complexity. *Perspect. Polit.* 7:65–70

Fang S. 2008. The informational role of international institutions and domestic politics. *Am. J. Polit. Sci.* 52:304–21

Fang S. 2010. The strategic use of international institutions in dispute settlement. *Q. J. Polit. Sci.* 5:107–31

Fang S, Stone RW. 2011. *International organizations as policy advisors*. Work. pap., Dep. Polit. Sci., Rice Univ.

Fearon JD. 1998. Bargaining, enforcement, and international cooperation. *Int. Organ.* 52:269–305

Gilligan MJ. 1997. *Empowering Exporters: Reciprocity, Delegation, and Collective Action in American Trade Policy*. Ann Arbor: Univ. Mich. Press

Gilligan MJ. 2004. Is there a broader–deeper trade-off in international multilateral Agreements? *Int. Organ.* 58:459–84

Gilligan MJ. 2006. Is enforcement necessary for effectiveness? A model of the international criminal regime. *Int. Organ.* 60:935–67

Gilligan MJ. 2009. The transactions cost approach to international cooperation. In *Power, Interdependence and Nonstate Actors in World Politics*, ed. HV Milner, Andrew Moravcsik, pp. 50–65. Princeton, NJ: Princeton Univ. Press

Gilligan M, Johns L, Rosendorff BP. 2010. Strengthening international courts and the early settlement of disputes. *J. Confl. Resolut.* 54:5–38

Gilligan TW, Krehbiel K. 1989. Asymmetric information and legislative rules with a heterogeneous committee. *Am. J. Polit. Sci.* 33:459–90

Gilpin R. 1981. *War and Change in World Politics*. New York: Cambridge Univ. Press

Greene WH. 2008. *Econometric Analysis*. Upper Saddle River, NJ: Pearson Prentice Hall

Grieco J. 1988. Anarchy and the limits of cooperation: a realist critique of the newest liberal institutionalism. *Int. Organ.* 42:485–507

Grossman GM, Helpman E. 1994. Protection for sale. *Am. Econ. Rev.* 84:833–50

Henkin L. 1968. *How Nations Behave: Law and Foreign Policy*. New York: Columbia Univ. Press

Hogg RV, McKean JW, Craig AT. 2005. *Introduction to Mathematical Statistics*. Upper Saddle River, NJ: Pearson Prentice Hall

Hollyer JR, Rosendorff BP. 2011. Why do authoritarian regimes sign the Convention Against Torture? Signaling, domestic politics, and non-compliance. *Q. J. Polit. Sci.* 6:275–327

Iida K. 1993. When and how do domestic constraints matter? Two-level games with uncertainty. *J. Confl. Resolut.* 37:403–26

Iida K. 1996. Involuntary defection in two-level games. *Public Choice* 89:283–303

Johns L. 2007. A servant of two masters: communication and the selection of international bureaucrats. *Int. Organ.* 61:245–75

Johns L. 2011a. Courts as coordinators: endogenous enforcement and jurisdiction in adjudication. *J. Confl. Resolut.* In press

Johns L. 2011b. *Depth versus rigidity in the design of international agreements*. Work. pap., Dep. Polit. Sci., Univ. Calif. Los Angeles

Johns L. 2011c. *Strengthening international courts: the hidden costs of legalization*. Unpublished manuscript, Dep. Polit. Sci., Univ. Calif. Los Angeles

Kennedy PM. 1987. *The Rise and Fall of Great Powers: Economic Change and Military Conflict From 1500 to 2000*. New York: Random House

Keohane RO. 1984. *After Hegemony: Cooperation and Discord in the World Political Economy*. Princeton, NJ: Princeton Univ. Press

Kindleberger CP. 1986. *The World in Depression, 1929–1939*. Berkeley: Univ. Calif. Press

Koremenos B. 2001. Loosening the ties that bind: a learning model of agreement flexibility. *Int. Organ.* 55:289–325

Koremenos B. 2005. Contracting around international uncertainty. *Am. Polit. Sci. Rev.* 99:549–65

Kydd A. 2001. Trust building, trust breaking: the dilemma of NATO enlargement. *Int. Organ.* 55:801–28

Kydd AH. 2010. Rationalist approaches to conflict prevention and resolution. *Annu. Rev. Polit. Sci.* 13:101–21

Mansfield ED, Milner HV, Rosendorff BP. 2002. Why democracies cooperate more: electoral control and international trade agreements. *Int. Organ.* 56:477–514

Mayer FW. 1992. Managing domestic differences in international negotiations: the strategic use of internal side-payments. *Int. Organ.* 46:793–818

McGillivray F, Smith A. 2000. Trust and cooperation through agent-specific punishments. *Int. Organ.* 54:809–24

McGillivray F, Smith A. 2004. The impact of leadership turnover on trading relations between states. *Int. Organ.* 58:567–600

McGillivray F, Smith A. 2006. Credibility in compliance and punishment: leader specific punishments and credibility. *J. Polit.* 68:248–58

McGillivray F, Smith A. 2008. *Punishing the Prince: A Theory of Interstate Relations, Political Institutions, and Leader Change*. Princeton, NJ: Princeton Univ. Press

Milgrom PR, North DC, Weingast BR. 1990. The role of institutions in the revival of trade: the law merchant, private judges, and the Champagne fairs. *Econ. Polit.* 2:1–23

Mo J. 1994. The logic of two-level games with endogenous domestic coalitions. *J. Confl. Resolut.* 38:402–22

Morrow JD. 1991. Electoral and congressional incentives and arms control. *J. Confl. Resolut.* 35:245–65

Morrow JD. 1994. Modeling the forms of international cooperation: distribution versus information. *Int. Organ.* 48:387–423

Nalebuff B. 1991. Rational deterrence in an imperfect world. *World Polit.* 43:313–35

Niou E, Ordeshook P. 1994. Less filling, tastes great: the realist–neoliberal debate. *World Polit.* 46:209–34

Olson M. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard Univ. Press

Oye K. 1985. Explaining cooperation under anarchy: hypotheses and strategies. *World Polit.* 38:1–24

Powell R. 1991. Absolute and relative gains in international relations theory. *Am. Polit. Sci. Rev.* 85:1303–20

Powell R. 1993. Guns, butter, and anarchy. *Am. Polit. Sci. Rev.* 87:115–32

Powell R. 1994. Review: anarchy in international relations theory: the neorealist–neoliberal debate. *Int. Organ.* 48:313–44

Putnam RD. 1988. Diplomacy and domestic politics: the logic of two-level games. *Int. Organ.* 42:427–60
Raustiala K, Victor D. 2004. The regime complex for plant genetic resources. *Int. Organ.* 58:277–309
Reinhardt E. 2000. Adjudication without enforcement in GATT disputes. *J. Confl. Resolut.* 45:174–95
Ritter EH, Wolford S. 2011. Bargaining and the effectiveness of international criminal regimes. *J. Theor. Polit.* In press
Romano CPR. 1998. The proliferation of international judicial bodies: the pieces of the puzzle. *NYU J. Int. Law Polit.* 31:709–51
Romano CPR. 2011. A taxonomy of international rule of law institutions. *J. Int. Dispute Settlement* 2:241–77
Rosendorff BP. 2005. Stability and rigidity: politics and the design of the WTO's dispute resolution procedure. *Am. Polit. Sci. Rev.* 99:389–400
Rosendorff BP, Milner HV. 2001. The optimal design of international institutions: uncertainty and escape. *Int. Organ.* 55:829–57
Ruggie JG. 1992. Multilateralism: anatomy of an institution. *Int. Organ.* 46:561–98
Schneider CJ, Urpelainen J. 2011. Accession rules for international institutions: a legitimacy–efficacy trade-off? *J. Confl. Resolut.* In press
Snidal D. 1985. The limits of hegemonic stability theory. *Int. Organ.* 39:579–614
Thompson A. 2006. Coercion through IOs: the Security Council and the logic of information transmission. *Int. Organ.* 60:1–34
Urpelainen J. 2011. Unilateral influence on international bureaucrats: an international delegation problem. *J. Confl. Resolut.* In press
Verdier D. 2008. Multilateralism, bilateralism, and exclusion in the nuclear proliferation regime. *Int. Organ.* 62:439–76
Voeten E. 2001. Outside options and the logic of Security Council action. *Am. Polit. Sci. Rev.* 95:845–58
Voeten E. 2005. The political origins of the UN Security Council's ability to legitimize the use of force. *Int. Organ.* 59:527–57
Wagner RH. 1983. The theory of games and the problem of international cooperation. *Am. Polit. Sci. Rev.* 77:330–46
Waltz K. 1979. *Theory of International Politics*. New York: McGraw-Hill
Waltz KN. 1954. *Man, the State and War: A Theoretical Analysis*. New York: Columbia Univ. Press
Young OR. 1979. *Compliance and Public Authority: A Theory with International Applications*. Baltimore, MD: Johns Hopkins Univ. Press