# JUSTICE, INSTITUTIONS, AND MULTIPLE EQUILIBRIA

by Roger B. Myerson, Economics Department, University of Chicago
http://home.uchicago.edu/~rmyerson/research/justice.pdf

Abstract. Schelling's concept of the focal-point effect in games with multiple equilibria is among the most important ideas in social theory. When justice is recognized as a criterion for identifying focal equilibria, we see how justice can affect the rational behavior of selfish economic actors. The foundations of political institutions can be understood in terms of focal equilibrium selection in a more fundamental game. This paper probes these ideas with some simple game-theoretic examples. Multiple equilibria are analyzed in a rival-claimants game, and this coordination game is extended to simple models of property rights, political institutions, boundaries, and economic investment.

## 1. Introduction

How can justice influence the decisions of a rational economic agent? In this paper, we seek a fundamental perspective on this question by considering some simple game-theoretic models where people's treatment of each other may be rationally guided by considerations of justice. In particular, this paper argues that the rational supply of justice is essentially connected with questions of multiple equilibria in games. More broadly, we argue here that the foundations of any social institution may be understood in terms of more fundamental games that have multiple equilibria.

Questions about the supply of justice are ancient in the literature of social philosophy. We may begin with a definition of justice cited in Book 1 of Plato's Republic: that justice is giving each person what is due to him. But it is difficult for economic theorists to accept Plato's subsequent suggestion that a good society could create an ample supply of justice simply by educating its rulers to love justice above all else. Such ideas of culturally-determined endogenous preferences are very hard to assimilate into economic analysis, because they hold the potential of trivializing all our questions about designing or reforming social institutions. It is an economist's job to develop methodologies for analyzing how changes in the structure of the

institution may affect people's behavior and welfare. In such questions, we do not want to assume that the problems of poverty could be solved simply by educating the poor to develop a taste for their plight. Virtually any institution could generate better social outcomes if everyone could be motivated solely by a benevolent desire to raise the aggregate welfare of the whole society. To avoid such trivialization of questions about institutional reforms, most economic analysis has been based on a restrictive assumption that individuals' fundamental motivations are generally selfish and materialistic. If an institution cannot perform well under such an assumption, then we have good grounds for seeking to reform it. The great successes of economic analysis have been based on this approach.

But if everybody is a selfish materialist, then how can anybody be motivated to treat another person better simply because such better treatment is justly "owed" to him? If there were only one decision-maker in the world, then a person who does not intrinsically care about behaving justly could be induced to respond to demands of justice only if he was otherwise completely indifferent among his behavioral options. Such complete indifference could not provide more than a fragile basis for the supply of justice. But when we consider games with many decision-makers, we often find multiple equilibria. That is, many different patterns of behavior among the players may be rationally sustained as unique best responses to each other. In such situations, criteria of justice may be a crucial determinant of each individual's rational behavior precisely because he expects everyone else's behavior to be influenced by the same concepts of justice. Thus, justice may be rationally supplied by selfish individuals because justice is a criterion for selecting among equilibria of a game.

Although this point is simple, it may have been relatively neglected in our literature because of a methodological bias against studying games that have many equilibria. When we find many equilibria, it seems that economic analysis cannot make firm predictions, which may be a bit embarrassing to professional economists. But I want to argue here that the supply of justice and the foundations of social institutions can only be understood in the context of games with multiple equilibria. In this regard, this paper is perhaps just an amplifying echo of Thomas Schelling's Strategy of Conflict (1960) which, in Chapter 3, introduced the idea of focal coordination in games with multiple equilibria. The main point of this paper may simply be the

observation that Schelling's focal-point effect needs to be understood as one of the great fundamental ideas of social philosophy.

The basic idea is a simple one. A Nash equilibrium is a possible prediction of behavior for all players in a game such that, if every player believes that the others will behave as predicted, then it is rational for each player himself for behave according to this prediction. Thus, any prediction which is not a Nash equilibrium cannot be rationally accepted by the players as an accurate description of what will happen in the game, because if everybody believed its accuracy as a way of predicting others' behavior then at least one player would want to violate the predictions himself. When we study a game that has only one equilibrium, this equilibrium must be the only rational prediction of how the players will behave, and so game theory seems very powerful. But games with many equilibria seem a bit more difficult, because this logic leaves us will many possible predictions. Nash (1953) considered a simple bargaining game that had an enormous multiplicity of equilibria, but he tried to get around this multiplicity by studying mathematical properties that would characterize a unique equilibrium in this set.

In Schelling's view, however, such multiplicity of equilibria was not an technical problem to be avoided, but was a fact of life to be appreciated. In Chapter 3 of Strategy of Conflict, he led the reader to consider a long list of games, each having many equilibria.

Let us consider one introductory example that is similar to many in Schelling's list. A group of players, all wearing name tags, are sitting in a circle. Each must independently write one player's name on a piece of paper. If they all write the same name then they all get $100, except that the person named gets $200. Otherwise they all get $0. The players have never met each other before. But just before they play, someone walks in, puts a big shiny crown on one player's head, and walks away. Everybody can see the person with the crown and can read his name tag. To get paid anything, the players need to coordinate; and they might well coordinate by all naming the person with the crown. This idea, once generally perceived, can become quite compelling. By focusing people's attention on the equilibrium where everyone names this player, the crown can create a social situation where everyone thinks that everyone will play according to this equilibrium, and so everyone finds it indeed in his own best interest to play according to this equilibrium. That is, the crown may function as a self-fulfilling prophecy that everyone will

name the player wearing it. The crown has no intrinsic effect on the payoff structure of the game, but it may affect each player's rational decision by affecting his expectations about what others will do.

So in a game with multiple equilibria, the indeterminateness of Nash equilibrium as a solution concept opens the door to other factors influencing the rational behavior of players. Anything that focuses everyone's attention on one equilibrium may lead the players to expect this equilibrium, and so to rationally fulfill it. These focal factors could be anything in the environment or heritage of the players, as long as they are all aware of it.

I suggested a crown in this story because crowns have functioned similarly in traditional political institutions. Political agents need to agree about who is their leader, and special crowns were used in ceremonial assemblies of medieval Europe to indicate the person who is generally recognized as national leader. In a nonEuropean culture where crowns have had no such meaning, but where important leaders have been marked instead by the possession of a special staff, giving such a staff to one of the players in this game might lead everyone to name him, even when another player came in with a fancy crown on his head. So details of cultural traditions may be decisive in games with multiple equilibria.

Such dependence of focal equilibria on the players' shared culture should not be seen as a failure of economic analysis. Instead, the role of culture as a focal factor should be recognized as a primary mechanism by which culture can influence rational economic behavior. People's understanding of justice is an essential part of their cultural tradition, and if rational economic behavior were independent of any cultural influences then justice would have no role in the determination of economic behavior.

But these focal factors can only work if they focus the players on an equilibrium. If we changed the payoffs in our first example so that each player instead gets $100 for each time that his own name is written by anybody in the game, then the unique equilibrium would be for each player to write his own name. In this case, a crown on one player's head, no matter how big and shiny, could not induce others to write his name instead their own.

2. Truthful equilibria in a simple reporting game

Before considering justice, let us start with truth. Consider a simple game where three players are each given a device with two buttons on it, one labeled "Yes" and the other labeled "No." Each player must independently decide which button to press. If he presses a button with the same label as at least one other player then he gets $100, but if he is the only person to push a button with this label then he gets $0. Each player wants only to maximize his expected monetary payoff.

This game has an equilibrium where they all choose Yes, but it has another equilibrium where they all choose No. The focal-point effect suggests that their behavior in this game may depend on payoff-irrelevant features of environment that the players all observe, including the broader cultural traditions that they share. So let us add that, before making their decisions in this game, the players together are shown a picture and are asked a question about it: "Is this a picture of Abraham Lincoln?"

Truthful reporting can be expected in this game when truth functions as the focal coordinating device. With truth as the focal factor, the players should all choose Yes if the picture is one that a reasonable person would believe to be Abraham Lincoln, but otherwise they should all choose No. Of course this focal understanding depends on many cultural conventions. The picture and question could not have such a focal coordinating role if the players did not know who Abraham Lincoln was, or if they did not understand the English language.

If the $100 payoff were changed to $1 or to $0.01 (but players remained solely motivated by the goal of maximizing expected monetary payoff), then our analysis of the truthful equilibrium would remain the same. In this sense, we find no lower bound on the cost of obtaining truth in this simple social system.

Real reporting systems involve more complex systems of cross checking and auditing. But the role of truth in much social discourse can be understood as a criterion for identifying the focal equilibrium in coordination games where people are penalized for making statements that are inconsistent with other testimony.

When there is a unique equilibrium, however, truth may lose this focal-coordination role. For example, suppose instead that each player would be paid $100 when his selection was

different from the other two players, but would get $0 if he matched one or more others. If there were no other sanctions against "false" reports then the role of truth would vanish in this revised game, where the unique equilibrium is to randomize between Yes and No with equal probability.

3. Justice from Basu's taxi driver and wars of attrition

To introduce justice, let us consider a simple question posed by Kaushik Basu (1990), in the introduction to his book on Analytical Development Economics, when he wanted to argue the importance of culture as a factor in economic development. Basu asked why a passenger should pay a taxi driver after getting out of the car, in a city where she is visiting for one day. One possible answer is that the driver is likely to fight with the passenger if she does not pay. But that answer raises the question of why the driver does not use the same threat to get a pedestrian to pay similarly even without having ridden in the taxi.

Some social theorists might answer that, with a good social education, drivers and passengers and pedestrians should all want to honor a Platonic definition of justice and give each person what is due. That is, passengers should have acquired a sense of duty to pay, but pedestrians feel no such duty, and each should want to do her duty. As discussed above, this endogenous-preference approach seems a poor foundation for economic analysis. If this were the only explanation, then we might conclude that justice is possible only when people are trained to deviate from their selfish materialist impulses. We need to know if there is some way that justice can be delivered by rational people even when they are selfish materialistic individuals.

Let's dismiss one other cheap answer: that the passenger should pay because a policeman might arrest her otherwise. We can dismiss this answer simply by assuming that the two people are alone on the street. But our reason for dismissing this answer is that, instead of simplifying the story, a policeman would really complicate it. Relying on a policeman for enforcement would beg the question of why the policeman should take the trouble of arresting a passenger who does not pay, but not arresting a pedestrian who refuses to pay. The policeman's motivations might depend on how his peers and superiors behave in the Police Department, which is a complex social institution. Complex social institutions are designed to provide justice

in various forms. But we may see the mechanisms of social justice more clearly if we limit ourselves at first to a society that includes only two people confronting each other alone.

Let us consider a simple game that formally models this situation. To avoid informational problems, let us assume that the passenger is known to have $20, which also happens to be the fare displayed on the taxi meter. To make the players' positions symmetrical, let us assume that the $20 has actually fallen out the passengers' pocket and is on the ground between them. Now the driver and the passenger each have a simple decision problem: to either claim the $20 or leave it and depart empty-handed. When a player leaves, his or her payoff is $0. If one player claims the money and the other player leaves, then the claiming player gets the $20 prize. If both players claim the money, then neither gets the prize and they suffer a cost of conflict worth some amount $c to each of them.

Let V = 20 denote the value of the monetary prize in question, and let us relabel the driver and passenger (or pedestrian) as players 1 and 2. Then this simple game of rival claimants can be written as follows:

| | Player 2 claims | Player 2 leaves |
|---|---|---|
| Player 1 claims | $-c, -c$ | V, 0 |
| Player 1 leaves | 0, V | 0, 0 |

**Table 1:** Payoffs to players 1 and 2 respectively in a simple rival-claimants game.

This game has three equilibria. There is one equilibrium in which player 1 (the driver) claims the prize and player 2 leaves empty-handed. Given that player 1 is expected to claim, leaving empty-handed is better for player 2 than suffering the cost of conflict ($0>-c$). When player 2 is expected to leave, player 1 should confidently claim the money ($V>0$). In this equilibrium, player 1's expected payoff is V=20, and player 2's expected payoff is 0.

There is a second equilibrium in which player 1 leaves and player 2 claims. Given that player 1 is expected to leave, player 2 should confidently claim the prize. When player 2 is expected to claim, player 1 would be better off leaving to avoid the cost of conflict. In this equilibrium, player 1's expected payoff is 0, and player 2's expected payoff is V=20.

There is also a third equilibrium in which both players independently apply the same randomized strategy. To compute this randomized strategy, first consider player 1's decision problem, letting p denote the probability that player 2 will leave. If player 1 claims then his expected payoff is $Vp + -c(1-p)$, but if player 1 leaves then his payoff is 0. Notice that

$$Vp + -c(1-p) = 0 \text{ when } p = c/(V+c).$$

Thus, player 1 is willing to randomize when player 2 is expected to leave with probability $c/(V+c)$, because player 1's expected payoff is then \$0 no matter what player 2 does about the money. Similarly, player 2 is willing to randomize when player 1 leaves with probability $c/(V+c)$. Thus, there is a symmetric randomized equilibrium in which each player independently claims with probability $V/(V+c)$ or leaves with probability $c/(V+c)$. In this randomized equilibrium, each player's expected payoff is 0.

The assumption that the players must decide, once and for all, whether to claim the prize forever or leave immediately may seem unrealistic. In real life, conflicts are dynamic open-ended processes where each party can regularly reconsider the question of whether to quit or keep fighting. So let us now let us make the game a bit more realistic by allowing that, if both players initially claim the prize, then each of them can regularly reconsider the option of leaving after any each interval of conflict, say once every minute. In this dynamic context, the parameter c should be reinterpreted as only the cost of struggling with the other player for a minute. In a conflict for V=\$20, the cost of arguing for one minute might be c=\$1 or less; so let us think of c as a small positive number, much less than V

$$V > c > 0.$$

In this dynamic version of the game, we allow that the players may both continue to claim the money for any number of minutes, but they will each pay the cost c for every minute that their conflict continues. This dynamic game ends only when somebody finally decides to leave empty-handed, and then the other player will get the \$V prize, unless they both decide to leave at the same moment. (In equilibrium, the probability of such simultaneous departures will be small.)

For each of the three equilibria that we found in the one-stage game, there is a sequential equilibrium of this dynamic game in which the expected behavior is the same at each one-minute

round until somebody leaves. There is one equilibrium, with expected payoffs (V,0), in which player 1 is always expected to claim and player 2 is always expected to leave in the next minute, no matter how long they may have both tried to claim in the past. Similarly, there is a second equilibrium, with expected payoffs (0,V), in which player 2 is always expected to claim and player 1 is always expected to leave in the next minute. In either case, when one player is expected to always persevere, it is rational for the other player to leave as soon as possible; and a rational player should indeed persevere when his or her opponent is expected to surrender soon.

The dynamic game also has a sequential equilibrium in which each player independently applies the same randomized strategy, leaving with probability $c/(V+c)$, every minute until somebody finally leaves. So the length of conflict is random in this equilibrium. To verify that this scenario is a sequential equilibrium, consider the position of one player after they have been in conflict for some T minutes. If this player leaves now then his payoff will be $-cT$, where $cT$ is his sunk cost of past conflict. But if this player claims now for just one more minute, then his payoff after another minute will be either $V-cT$ or $-c(T+1)$, depending on whether the other player leaves or not in the next minute. The probability of the other player leaving in the next minute is $c/(V+c)$ in this equilibrium. Notice

$$(V-cT)[c/(V+c)] - c(T+1)[V/(V+c)] = -cT.$$

So at every round, the player is always indifferent between leaving immediately and continuing the conflict for one more minute, and so he is indeed willing to randomize. In particular, each player could get his expected equilibrium payoff by leaving immediately at the first minute, and so his expected payoff in this equilibrium is 0. Thus, we have a symmetric randomized sequential equilibrium in which both players get expected payoff 0. The expected duration of the conflict is just long enough to exactly cancel out the expected gains from possibly winning the \$V prize. This equilibrium is called a war of attrition.

In the war-of-attrition equilibrium with V=20 and c=1, each player's probability of leaving at the next minute is always $c/(V+c) = 0.048$ as long as the conflict continues. The expected duration of conflict in this equilibrium is 9.76 minutes, and the probability of nobody getting the prize because of simultaneous decisions to leave is only 0.024.

The above three equilibria have the property that they are stationary, in the sense that the

expected behavior stays the same no matter how long the conflict continues. There are other nonstationary sequential equilibria that may be worth noting. In particular, consider an assumption that, if both players were to claim at the first round of the game then, after the first minute of the game, they both would be expected to play according to the randomized war-of-attrition equilibrium at all subsequent rounds of the game. Because the players' expected payoffs are both 0 in the war of attrition, predicting such behavior in the subgame after the first minute will make the players' expected payoffs in the whole dynamic game just equal to the payoffs that they get in the first minute. So the three one-stage equilibria that we found for Table 1 (having expected payoff allocations (V,0) and (0,V) and (0,0)) can also be extended to sequential equilibria of the dynamic game with this equilibrium assumption, that a war of attrition will be played at all subsequent rounds if nobody leaves in the first one-minute round.

From our perspective, this multiplicity of equilibria should not be seen as an embarrassing technical difficulty, but rather as just what we need to get the "right" answer both for Basu's driver and passenger on one street, and for the economically-identical driver and pedestrian on a neighboring street. When sunk costs and unchangeable past events are ignored, the economic structure of decision-options and payoffs is identical in both situations. If the game had only one equilibrium then we would have to predict the same outcome in both situations. The multiplicity of equilibria allows us to suppose that, on one street, the driver and passenger will focus on the equilibrium in which the passenger immediately pays, because they both understand that the $20 rightfully belongs to the driver. But on the other street, where a similar driver sees a pedestrian, who has not ridden in the taxi but has similarly just dropped $20 on the ground, they both understand that the $20 rightfully belongs to the pedestrian. So the driver and pedestrian focus on the equilibrium in which the pedestrian "rightfully" claims her $20 and the driver leaves empty-handed. Thus, justice can be rationally supplied in our simple two-person society because the players' shared understanding of justice may be the focal factor that determines the equilibrium that they play, even though each player is a selfish materialist with no intrinsic preference for justice.

Good cultural education may be important here, not because it taught anyone to love justice above his or her own self interest, but because it taught them all to expect justice. In this

society, each individual is driven to give justice by the others' expectation that justice will be given to them. Where a shared cultural heritage has created a common understanding of what is owed to each individual, then the imperative to let everyone have what is owed to him may be derived from its being the focal equilibrium in a game of multiple equilibria.

## 4. A fable about the foundation of institutions

Let us now extend the preceding example to create a simple game-theoretic model of the origins of political institutions. Imagine an island with a large population of individuals. Every morning, these islanders may assemble in various groupings to talk and watch the sun rise. Then the islanders scatter for the day. During the day, the islanders are randomly matched into pairs who meet at random locations around the island, and each of these matched pairs plays a rival-claimants game, as in Table 1 above, with a given prize worth $V>0$ and a given cost of conflict $c>0$. Another round of this process is repeated every day. Each player's objective is to maximize the present discounted total value of his or her sequence of payoffs from these daily rival-claimants matches, with some given discount factor $\delta$ that is between 0 and 1.

One long-run equilibrium of this process is for everyone to play the symmetric randomized equilibrium in his match each day. This equilibrium represents an anarchic state of nature, where the value of every resource is dissipated in conflict, and so everybody's expected payoff is always 0. But rising up from this primitive anarchy, the players could develop cultural expectations which break the symmetry among the matched players, so that they will share an understanding of who should claim the prize and who should leave empty-handed.

One possibility is that the islanders might develop a traditional understanding that each player has a special "ownership" relationship with some region of the island, such that a player is expected to claim whenever he is in the region that he "owns". Notice that this system of ownership rights is a self-enforcing equilibrium, because the other player does better by leaving empty-handed (getting 0 rather than $-c$) when he expects the "owner" to take his claim, and so the owner should indeed claim the prize confidently. But such a system of traditional claiming rights might fail to cover many matching situations where no one has clear ownership. To avoid

the wasteful symmetric equilibrium in such cases, other ways of breaking the players' symmetry may be needed.

It might be possible to define more general principles that determine claiming rights in more situations, but these principles will be effective only if they are commonly understood by everybody. So the islanders might meet together to discuss and ratify new principles to resolve more rivalries. Thus the role of an assembly with legislative powers can be introduced here.

In cases that are not covered by any such generally understood laws, a pair of matched players might instead invite some third player to act as an arbitrator. Even though the arbitrator has no direct means to compel either player to use any strategy, if the arbitrator recommends one of the equilibria then his recommendation can become a self-fulfilling prophecy, as long as each player believes that the other will play according to the focal arbitrator's recommendation.

An arbitrator who wants to be fair could offer each player a positive expected payoff by basing his recommendation on some random event like the toss of a coin. He might recommend that player 1 should claim and player 2 should leave if the coin is Heads, but that player 2 should claim and player 1 should leave if the coin is Tails. Such random arbitration has the advantage that it offers a positive expected payoff to both players before the coin it tossed. It will not work, however, if the player who is disappointed after the coin toss would demand that another arbitrator should toss another coin. So for the players to be coordinated by a random device, they need some way to focus themselves on one randomization that cannot be repeated or appealed.

To focus attention with no higher appeal, it would be best to consult the highest possible authority. If the players share a cultural understanding that certain unpredictable processes may be used by the fundamental Spirit of the universe to answer questions, and that this Spirit will not allow itself to be bothered about the same question more than once, then a recommendation that is based on such a sacred randomization can serve as a focal coordination device that cannot be appealed to any higher arbitrator. The oracle's recommendations can be self-enforcing without any further intervention by the Spirit, provided that the recommendations to the players form an equilibrium. Thus our model can admit an important role of oracles and divination as an effective foundation for social coordination. (For example, see Evans-Pritchard, 1937. Moore, 1957, also offers game-theoretic perspective on divination.)

To avoid disagreements about who is to provide focal arbitration in a case, the islanders could develop a generally agreed system of leadership. That is, the islanders might appoint one of their population to serve as a leader, and this leader could announce each morning a set of instructions that specify which one of the two players should be claiming the prize in each of the daily matches. As long as the leader's instructions are clear and comprehensive, the understanding that every player will obey these instructions is a self-enforcing equilibrium. A player who claimed when he was expected to leave would only lower his expected payoff from 0 to $-c$, given that the other player is expected to take the prize here, as designated by the leader.

To make this system of government work on our island, the islanders only need a shared understanding as to who is the leader. Rituals in which the fundamental Spirit of the universe is seen to sanctify the leader's authority might also help to achieve such a focal consensus. But almost any imaginable criterion or process could be used to identify or select the leader. The leader might be the eldest among the islanders, or the tallest, or the one with the loudest voice. Or the islanders might determine their leader by some contest, such as a chess tournament, or by an annual election in which all the islanders vote. The islanders could use any method of selection that they can understand, because everyone will want to obey the selected leader's equilibrium instructions as long as everyone else is expected to obey him. Thus, self-enforcing rules for a political system can be constructed arbitrarily from the equilibrium-selection problem in this game.

The islanders could impose limits to a leader's authority in this political system. For example, there might be one leader whose instructions are obeyed on the northern half of the island, and another leader whose instructions are obeyed on the southern half. The islanders may even have ways to remove a leader, such as when he loses some re-election contest or when he announces arbitration decisions that violate some perceived limits. If a former leader tried to make an announcement in his former domain of authority, every player would be expected to ignore this announcement as irrelevant cheap talk.

Mathematical models in social science are like fables or myths that we read to get insights into the social world in which we live. Of course, the real world is very different from the simple island of this fable. But as in this island, coordination games with multiple equilibria are

pervasive in any real society. The basic point is that any successful society must develop leadership structures that can coordinate people's expectations in situations of multiple equilibria. So the first point of this fable is the basic social need for leadership and for political institutions that can provide it. (See also Calvert, 1995, pp 241-244 for articulate discussion of this point.)

Suppose that the players' payoffs in our model can also be interpreted as resources that increase their long-term reproductive fitness. Then an anarchic island where these resources are wasted in the symmetric equilibrium would sustain a much smaller population than another island where the players have systems of authority to coordinate them on better equilibria. If players from highly populated islands can colonize underpopulated islands, taking with them their cultural system of focal-equilibrium selection, then an archipelago of such islands should eventually be inhabited only by people who have systems of authority to coordinate them in matches where their are multiple equilibria. Thus, in any cultural tradition that has survived into the modern world, we should expect to find generally-accepted systems of rights and authority that provide effective focal coordination in most of the important games with multiple equilibria that may arise in daily life. In this regard, experimental games that are conducted in laboratories may be systematically different from real life. When we study real institutions, we may expect multiple-equilibrium problems to be resolved by existing social traditions of authority and culture, rather than by some abstract mathematical properties of the equilibria themselves.

The second point of this fable is that the effectiveness of a political institution may be derived simply from a shared understanding that it is in effect. The remark that our islanders might choose their leader by a chess tournament is meant to suggest that the rules of any higher-order social institution (here all the rules of a chess tournament) may itself be sustained as an equilibrium in a broader and more fundamental game that has an enormous multiplicity of equilibria. Applied social theorists must understand that there are games within games in the real world, because the selection of a focal equilibrium in a big game can effectively define an institution which itself can be the subject of game-theoretic analysis. As Hardin (1989) has argued, any political system may be understood as one of many possible equilibria of a more fundamental coordination game of constitutional selection. To the extent that political leaders can create general rules and guidelines for the creation of new social and economic institutions

throughout society, the political process of selecting a constitution may be viewed as the equilibrium-selection problem to solve all other equilibrium-selection problems in society.

A third point of this fable is that norms of justice may be largely sustained by systems of lower-order interactions among small groups of people. Although resources could be rationally expended in paying leaders or in contests for leadership, our model also admits equilibria where justice is provided at no cost in every dyadic interaction. Any player can be costlessly motivated to leave, when justice demands it, by the simple expectation that his opponent will justly claim the prize. Of course, the dyadic basis of enforcement here is due to the main simplifying assumption of this model: that the entire social system is built up from elementary interactions among pairs of players whose coordination problem would admit multiple equilibria even in isolation from all other interactions. In real life, the elementary coordination problems may involve larger numbers of people. But as in this model, the main deterrence to blatant thievery may indeed be the difficulty of claiming things that other people expect to be left for them. Thus, this model may be considered as an explanation of how there can be so little need for adjudication in most daily social interactions.

Any system of equilibrium expectations about claiming rights in this model can be considered as a system of justice for our islanders, although some systems of justice might seem unfair to an outside observer. In particular, although thievery may be effectively deterred in this simple story, bullying is quite possible in this model. There may be some bullies who are always expected to claim everything from everyone else, and there are equilibria in which everybody else acquiesces and leaves empty-handed in any interaction with such a bully. More generally, the islanders might develop a class system (or pecking order) such that, whenever players of different classes meet, the higher-class player always claims, confident of success, and the lower-class player always leaves empty-handed. If there is some initial doubt as to which of two groups will have the higher status, then a broader war-of-attrition equilibrium can exist in which members of these two groups continually claim against each other until somebody on one side backs down and causes his group to be understood as inferior in the class system.

5. Boundaries and linkage between different areas of potential conflict

The concept of a boundary is fundamental to the structure of international relations, where even small infractions of a boundary typically evoke costly defensive responses. Schelling (1960) showed that his theory of focal equilibria may provide a basis for understanding the logic of such rigid boundaries. The central insight is that a player's failure to vigorously fight against even a small violation of a perceived boundary might lead others to believe that the player would also surrender much larger areas. This fundamental insight can be formalized by an extension of our rival-claimants game, where behavior in one match may influence focal equilibrium behavior in other subsequent matches.

For example, consider two players who will be matched every day to play many instances of our rival-claimants game (Table 1), but each of these matches will be located at a different place in a large field. We may imagine that the prizes in these matches are fruits that ripen each day on bushes scattered throughout the field. Suppose that player 1's home is north of the field, player 2's home is south of the field, and there is an old fence that crosses the field from east to west. The fence may be small, but it is the most conspicuous feature in the field. So the players may naturally focus on an equilibrium such that only player 1 is expected to claim in their matches north of the fence, and only player 2 is expected to claim in their matches south of the fence.

But suppose now that one bush has grown right through the fence, so that there is confusion about which side it is on. There may be an equilibrium in which both players claim in their matches here at the fence, even though the conflict will cost them both c each day. In this equilibrium, each player is willing to suffer the cost c each day because he anticipates that, if he ever failed to claim here at the boundary, then his opponent would expect to claim successfully everywhere. That is, in this equilibrium, if either player ever left the boundary prize uncontested, then their future behavior would switch to the equilibrium in which this player would be expected to leave all prizes uncontested everywhere, and so the other player would expect to claim successfully in all future matches on both sides of the fence. Thus, our model can explain a costly protracted conflict at a perceived boundary, as long as each player's expected cost of the conflict here is less than the expected total value of all his claiming rights.

6. Investment and the need for intertemporal linkage in equilibria

In the preceding sections, we have seen many pure-strategy equilibria that were equally efficient, in the sense of maximizing the expected sum of players' payoffs. Our analysis was greatly simplified by using a structure (in Table 1) in which any efficient outcome could be enforced as an equilibrium within each match by itself, without any expectation that any future behavior might depend what the players do in the current match. But implicit in the original taxi-driver story was that, to sustain a taxi industry, the players' focal equilibria in rival-claimants matches must depend on whether the taxi driver has actually given the other player a ride. In this section, we augment our basic model to make this idea explicit. The result is a model that shows how efficient equilibria may require that focal equilibrium selection in different subgames should depend on the players' past behavior. In these efficient equilibria, sunk costs must matter.

So let us now augment the basic rival-claimants matches (from Table 1) by adding two preliminary stages. At the first preliminary stage, player 1 can select any investment expense $x \geq 0$, where more investment will improve the prize. At the second preliminary stage, after observing this investment x, player 2 selects any payment $y \geq 0$ to give to player 1. Finally, knowing both x and y, players 1 and 2 play a version of the rival-claimants game. But let us now assume that the value of the prize to each player i is some function $V_i(x)$ that depends on player 1's initial investment. Thus, the players' payoffs now depend on whether each leaves or claims the prize as in Table 2.

| | Player 2 claims | Player 2 leaves |
|---|---|---|
| Player 1 claims | $y-x-c, -y-c$ | $V_1(x)+y-x, -y$ |
| Player 1 leaves | $y-x, V_2(x)-y$ | $y-x, -y$ |

**Table 2:** Payoffs in a rival-claimants game after 1's investment x and 2's payment y.

Given any values of x and y from the preliminary stages of this game, the three strategy-pairs that were equilibria in Table 1 are also equilibria in the third-stage subgame here in Table 2. But in an equilibrium of the overall game, the selection among these three equilibria could

depend on the prior selections of x and y. For example, they might expect player 2 to claim the prize if $y \geq Y(x)$ but player 1 to claim the prize if $y < Y(x)$, where $Y(x)$ is some function that can be interpreted as the "just price" for this prize with the investment x.

Let us assume that $V_2(x) > V_1(x) > 0$ for all x, and that the maximum of $V_2(x) - x$ is achieved at some positive investment $\bar{x} > 0$; that is,

$$\max_{x \geq 0} V_2(x) - x = V_2(\bar{x}) - \bar{x}.$$

So total-payoff maximization would require that player 1 should invest this amount $\bar{x}$ and player 2 should get the prize. There are some equilibria of this three-stage game in which this efficient economic transaction is accomplished. For example, total-payoff maximization could be achieved if the players' focal-equilibrium expectations about who should claim the prize in the final stage would be determined by comparing player 2's second-stage payment to a just-price function of the form $Y(x) = V_2(x) - K$, where K is some constant such that $0 \leq K \leq V_2(\bar{x}) - \bar{x}$. With the highest possible K, player 2 would just pay $\bar{x}$ when player 1 invested this amount, and player 2's expected equilibrium payoff would be $V_2(\bar{x}) - \bar{x}$.

But any efficient equilibrium here requires a nontrivial intertemporal linkage in the determination of expected equilibrium behavior for the final subgame. If player 2 would be always expected to claim the prize, regardless of the prior investment and payment (x,y), then player 1 would rationally choose the minimal investment x=0.

This example also highlights the dynamic consistency problem that confronts a leader who has the power to select the focal equilibrium in any subgame. If player 2 is a strong leader who will be able to select the focal equilibrium in the Table-2 subgame after x and y have been chosen, then player 2 will always select the equilibrium where 2 claims and 1 leaves. Anticipating such expropriation by the leader, player 1 should not make any positive investment, and so player 2's expected payoff will be only $V_2(0)$, which is less than $V_2(\bar{x}) - \bar{x}$. Thus, a leader may be better off if he loses his ability to influence people's equilibrium expectations after the beginning of the game. Even an absolute monarch may want to impose limits on his own future authority.

## 7. Conclusions

We have argued that truth and justice can be understood game-theoretically as criteria for identifying a focal equilibrium in games with multiple equilibria. The cost of obtaining justice in society is indeterminate in our analysis, however. In our island of rival claimants, we saw that leaders might judge their own cases more favorably, which would be a cost for everyone else; but if everybody could perfectly observe the leader's actions then there would also exist equilibria where such a self-serving leader would be immediately replaced. When leaders can be only imperfectly observed by their followers, however, then we may find some necessary positive cost of inducing them to properly administer justice. Such a model has been considered by Milgrom, North, and Weingast (1990), but in their analysis of games with potentially-dishonest judges they found that incentives for honest adjudication could be provided without paying judges much more than the cost of their time. So more work is needed to develop models that might provide testable predictions about the cost of reliable adjudication.

The questions of justice that we considered in our models here have been generally about property rights rather than contractual enforcement. Contractual disputes involve failures of trust between parties who have entered into some kind of relationship. But concepts of reputation, trust, and other kinds of interpersonal relationships can be understood game-theoretically in terms of multiple equilibria in repeated games (such as the repeated prisoners' dilemma games that Milgrom, North, and Weingast studied). In such repeated games, a pattern of mutually-beneficial generosity among players may be sustained as an equilibrium by a threat that, if any player ever deviated from this pattern of cooperative behavior then everyone's expected behavior would switch to another equilibrium in which the deviator's expected payoff would be worse. That is, a good equilibrium in which players trust each other may depend on possibility of switching to other equilibria in which the players do not trust each other. Thus, relationships of trust and distrust can be understood as different equilibria in a repeated game.

REFERENCES

Kaushik Basu, Analytical Development Economics, MIT Press (1997).

Randall L. Calvert, "The Rational Choice Theory of Social Institutions: Cooperation, Coordination, and Communication," in Modern Political Economy, edited by Jeffrey S. Banks and Eric A. Hanushek, Cambridge U. Press (1995), pp. 216-267.

E. E. Evans-Pritchard, Witchcraft, Oracles, and Magic among the Azande, Oxford: Clarendon Press (1937).

Russell Hardin, "Why a Constitution," in The Federalist Papers and the New Institutionalism, edited by B. Grofman and D. Wittman, NY: Agathon Press (1989), pp. 100-120.

Paul R. Milgrom, Douglass C. North, and Barry R. Weingast, "The Role of Institutions in the Revival of Trade: the Law Merchant, Private Judges, and Champagne Fairs," Economics and Politics 2 (1990), 1-23.

Omar K. Moore, "Divination – A New Perspective," American Anthropologist 59 (1957), 69-74.

John F. Nash, "Two-Person Cooperative Games," Econometrica 21 (1953), 128-140.

Plato, The Republic, edited by G. R. F. Ferrari, translated by Tom Griffith, Cambridge U. Press (2000).

Thomas C. Schelling, The Strategy of Conflict, Harvard U. Press (1960).

*This paper is to be published in Chicago Journal of International Law 5 (Summer 2004).*