# Game Theory and Institutions

Ken Binmore
Economics Department
Bristol University
8 Woodland Road
Bristol BS8 1TB, UK

**Abstract:** This short paper begins with a summary of the views of a sympathetic game theorist on the current state of play in what is still called the New Institutional Economics. It continues with a much abbreviated summary of my own attempts to treat justice as a kind of institution in the hope that this will serve as a case study in how game theory can serve as a useful intellectual framework for the study of human institutions.

# Game Theory and Institutions[1]

**by Ken Binmore**

## 1 Institutional Economics

This short paper begins with a summary of the views of a sympathetic game theorist on the current state of play in what is still called the New Institutional Economics. It continues with a much abbreviated summary of my own attempts to treat justice as a kind of institution in the hope that this will serve as a case study in how game theory can serve as a useful intellectual framework for the study of human institutions.

**What's new?** Working economists like myself have little or no idea of what went on in the old institutional economics. Whatever institutional economists did before Coase and Williamson has now been so eclipsed by modern developments that it now seems unnecessary to distinguish between the old institutional economics and the new. So why do institutional economists persist on calling their subject the New Institutional Economics rather than just Institutional Economics?

I think that part of the reason is to signal an attachment to various pieces of baggage that were useful in the early history of the subject, but now stand in the way of the kind of progress that would be possible by integrating institutional economics with game theory. We are aready more than halfway there, as indicated by the following list of topics said by Wikipedia to typify the New Institutional Economics:

> Organizational arrangements, transaction costs, credible commitments, modes of governance, persuasive abilities, social norms, ideological values, decisive perceptions, gained control, enforcement mechanism, asset specificity, human assets, social capital, asymmetric information, strategic behavior, bounded rationality, opportunism, adverse selection, moral hazard, contractual safeguards, surrounding uncertainty, monitoring costs, incentives to collude, hierarchical structures, bargaining strength.

With two or three exceptions, these are all topics pursued vigorously within the game theory community.

**Transaction costs.** I want to draw particular attention to two pieces of baggage. The first is Oliver Williamson's [14] notion that the world is best seen as a distortion from the ideal of perfect competition created by the existence of transaction costs.

It is clear why the transaction-costs paradigm was a useful rhetorical device at the time Williamson wrote, and nobody would want to deny that transaction costs

[1]I am grateful to the British Arts and Humanities Council for funding this work through grant AH/F017502/1.

matter a great deal. However, the world has moved on since the seventies. We now understand that transaction costs are only one of numerous factors that shape imperfectly competitive markets. More important factors are commonly the unregulated exercise of market power and the exploitation of asymmetries of information.[2]

I think we need to re-orientate ourselves, seeing imperfect competition as the norm, both within and between institutions. Perfect competition would then be displaced from its central role, and seen rather as a limiting case of imperfect competition in which all the many different kinds of friction have become negligible.

The benefit from adopting the paradigm shift would be a sharpening in the modeling of particular institutions as a result of bringing the methods of game theory to bear in a more systematic way than at present. But there would also be a cost. Tolstoy famously says that all happy families are the same, but each unhappy family is unhappy in its own way. Similarly, all perfectly competitive markets are the same, but each imperfectly competitive market is imperfectly competitive in its own special way. General principles are therefore unlikely to survive the paradigm shift. Personally, I think little would be lost thereby. For example, attempts to regulate imperfectly competitive industries based on general principles derived from analogies with perfect competition or pure monopoly are notoriously unsuccessful. There really is no alternative to spending serious money researching the unique features of each industry needing to be regulated and tailoring the regulation to the special circumstances of the industry.

**Rules of the Game?** Douglass North [10] has been very influential in arguing that an institution should be characterized in terms of the rules that govern its operation. These rules of the game are to be understood as including, not only formal legal rules, but the informal social norms that govern individual behavior and structure social interaction within the institution.

For such "rules of the game" to correspond to rules of the game as understood in game theory, they need to be unbreakable. That is to say, the players must not have strategies available to them that can result in a rule being violated. Can a social norm have such a character? The answer depends on the circumstances.

For example, it may be that the punishment that would be inflicted by one's fellows for falling foul of a social norm may be so severe that nobody would ever contemplate taking the risk while in their right mind. In modeling the institution, one might then sensibly specify obedience to the social norm as being one of the rules of the game. By doing so, one loses the opportunity to study the social norm itself, since its characteristics are imposed on the model by fiat, but studying the reasons for the evolution and survival of this particular social norm may not be the purpose of the analysis.

For example, when considering how best to regulate an industry, one may write

[2]The concept of a transaction cost is sometimes stretched to include all factors that result in a deviation from perfect competition, but such a fix at the conceptual level does not help when real phenomena need to be modeled.

a model in which the agents are assumed to honor the requirements of the criminal law. Of course, we can all cheat and steal if we want to, but if the monitoring and enforcement mechanisms put in place by society are sufficiently effective, then nobody will actually cheat or steal, and so the result is the same as if the agents were literally incapable of cheating and stealing. In brief, when adequate enforcement mechanisms are in place, then it makes good sense to follow North's advice and to treat social norms (like observing the law or maintaining political correctness) as though they were as strictly binding on the agents as the rules of a formal game.

However, it is commonplace for models to be constructed in which the need to locate an actual enforcement mechanism in real life to justify assuming unbreakable rules in a model is not even considered. The most flagrant examples arise in constitutional studies when appeals are made to the theory of mechanism design in proposing ideal constitutions.[3] But there is no external enforcement agency by definition when we are talking about the whole constitution of an independent nation. The rules written into the constitution must therefore be self-enforcing if they are to operate successfully.

One only has to compare the constitution of the USA with that of the old USSR to see why. The latter document is a wondrous list of utopian aspirations, but entirely useless to the citizens of the USSR because those holding power held its provisions in contempt. The constitution of the USA, on the other hand, was written by people who understood that the best a constitution can do is to help in coordinating behavior on one of the many equilibria in the actual game of life that would be played by American citizens in the future. Accepting the pretence that the Supreme Court merely reinterprets consitutional provisions rather than rewriting them to bring them up to date with current opinion, its survival is a tribute to the hard-headed realism of its authors.

In brief, a constitution is not a substitute for the unbridled exercise of power. One cannot make power go away by writing words on a piece of paper. The best a constitution can do is to focus attention on a particular way of balancing power. To overlook this point is to fail to understand constitutional issues completely.

The game theory solution to the problem that arises when the rules of an institution are not enforced by some incorruptible external agency is to move to a larger game—the game of life—in which the institution is regarded as being embedded. This larger game must have rules that are genuinely unbreakable, like the laws of physics, because if the players had strategies whose implementation resulted in the rules being broken, then they would not be the genuine rules of the game. The rules of the institution to be studied then have a lesser status. The players can break them if they want to, but if the institution is stable, the rules will not be broken, because obeying them is part of the behavior required by an equilibrium of the game. That is to say, an institution is not treated as a game itself, but as part of the description of an equilibrium within a larger game of life.

[3]The word $mechanism$ in this context refers to those rules of a game that do not relate to the personal preferences and beliefs of individual agents.

## 2 Equilibrium Selection Problem

The rest of this paper outlines my theory of justice with a view to offering a case study of how an institution can be seen as an equilibrium in an unvarying game of life (Binmore [4, 3, 2]). I am interested in the institutions that collectively determine the social contracts of societies—especially the ancestral hunter-gatherer societies within which our sense of fair play presumably first evolved. I think that we evolved the capacity to entertain fairness norms because they allowed our species a quick and efficient way to solve the coordination problems that inevitably arise when a group is faced with a new situation. For example, how should a novel source of food be shared without fighting or other wasteful conflict? If I am right, then fairness can be seen as evolution's solution to the equilibrium selection problem that arises in certain games with multiple equilibria.

The idea that fairness arose as a solution to an equilibrium selection problem sometimes creates confusion because critics do not know whether to classify the theory as neoclassical or behavioral. In fact, it belongs to neither category. It differs from the tradition in neoclassical economics in recognizing that almost any realistic game has many Nash equilibria. Even perfectly competitive markets have multiple Nash equilibria if they are repeated every day. It then ceases to be true that there is a necessary trade-off between equity and efficiency. Neoclassical arguments to this effect rely on models that have only one equilibrium. More generally, I think that it is only because traditional neoclassical economics largely confined its attention to models with a unique equilibrium that it managed to get so far without paying any serious attention either to fairness norms or to all the many other institutions without which our society would fall apart.

My theory differs from the standard behavioral approach in not treating fairness as a property of utility functions. People doubtless do have other-regarding preferences to a greater or lesser degree. How else are we to explain why most of us give money to charity? However, I think that a theory of fairness which says that people play fair because they like playing fair is painfully naive. Nor am I satisfied that the empirical support claimed for this approach from experimental work is anywhere near adequate (Binmore and Shaked [5]). In my theory, individuals may or may not have other-regarding preferences, but even if they were all entirely selfish, they would still be stupid to ignore the fairness norms that operate in their society.[4]

**Social contracts.** A social contract is the set of common understandings that allow the citizens of a society to coordinate their efforts.

The common understandings or conventions that make up a social contract are many and various. They range from the arcane table manners we employ at formal dinner parties to the significance we attach to the green pieces of paper we

[4]The behavioral literature often blurs the distinction between a social preference and a social norm, but it is important in my work to make a sharp distinction between a payoff function and an equilibrium selection device.

carry round in our wallets bearing pictures of past presidents. From the rules that govern how we drive our cars in heavy traffic to the meaning of the words in the language we speak. From dietary and sexual taboos to the standards of integrity and truthfulness expected of honorable folk. From the vagaries of fashion to the criteria that we fondly suppose secure ownership of our possessions. From the amount that we think it appropriate to tip in restaurants to the circumstances under which we are ready to submit ourselves to the authority of others.

The seemingly profound half of each of these pairings is usually attributed to iron laws of morality derived from an ineffable source into which we are not encouraged to look too closely. But I think the differences between the trivial and the profound halves in the pairings are differences only of degree. There are no iron laws beyond those encoded in our genes. Like Gulliver in Lilliput, we are bound only by a thousand gossamer threads woven from our own beliefs and opinions.

It is true that some of our cultural conventions are codified as laws, but the legal system and the constitution of a modern society are relevant to a social contract only to the extent that they are actually honored in practice. If it is customary to give and take bribes, then giving and taking bribes is part of the social contract, whatever the law may say. Nor are popes, presidents, kings, judges, or the police exempt from the social contract of the society in which they officiate. Far from enforcing the social contract, they derive what power they have from a social convention which says that ordinary citizens should accept their direction. If they were ignored in the same way that the citizens of Naples ignore traffic signals, they would be totally powerless.

**How do social contracts work?** What is the glue that holds a society together? It isn't the law or the constitution. These are just words on a piece of paper. It isn't the officers of state. They are just people like you or me. Is it our sense of moral obligation? But there is honor of a sort even among thieves. Is it God? Some social contracts are so horrendous that even fundamentalists must sometimes entertain doubts.

None of these answers fit the bill, because we asked the wrong question. A stable social contract doesn't need any glue. People follow its precepts, because they wil be rewarded if they do and punished if they do not.

When punishment is mentioned, one's mind naturally turns to electric chairs and torture chambers. But the punishments that deter us from cheating on the social contract are nearly always so mild that we scarcely notice them at all. In my own country, I find it hard to codify the subtle use of body language and shades of verbal expression that my neighbors use to hint that my current behavior is likely to result in more positive forms of social exclusion if I do not start mending my ways. One is so habituated to responding appropriately, that such subliminal signals are automatically translated into behavior without any conscious control. Only when some social gaffe in a foreign country is greeted by an unfamiliar signal does the mechanism become apparent.

In David Hume's metaphor, a social contract holds together like a drystone wall or a masonry arch. Each stone supports and is supported by its neighbors, without any need for cement or glue. In modern game theory, we express the same idea by saying that the rules of a stable social contract succeed in coordinating our behavior on an *equilibrium* in the game of life.

So what is an equilibrium? How does it capture Hume's big idea? How come a society finds itself at one equilibrium rather than another?

**Equilibrium.** A game is any situation in which people or animals interact. The plans of action of the players are called strategies. A Nash equilibrium is any profile of strategies—one for each player—in which each player's strategy is a best reply to the strategies of the other players.

A very simple example is the Driving Game we play each time we get into our cars in the morning to drive to work. Shall we drive on the left or on the right? If all we care about is avoiding accidents, the game has three Nash equilibria. In the first, we all choose the strategy of driving on the left. In the second, we all choose the strategy of driving on the right. In the third, we each toss a coin to decide whether to drive on the left or on the right. The third alternative may seem dubious, but if everybody else is randomizing their choice, your chances of ending up in an accident are going to be the same whatever you do. So tossing a coin is just as much a best reply as doing anything else.

Why should anyone care about Nash equilibria? There are at least two reasons. The first is that if a game has a rational solution that is common knowledge among the players, then it must be an equilibrium. If it were not, then some of the players would have to believe that it is rational for them not to make their best reply to what they know the other players are going to do. But it can't be rational not to play optimally.

The second reason why equilibria matter is even more important. If the payoffs in a game correspond to how fit the players are, then evolutionary processes—either cultural or biological—that favor the more fit at the expense of the less fit will stop working when we get to an equilibrium, because all the survivors will then be as fit as it is possible to be in the circumstances.

**Reciprocal altruism.** The invisible hand can only be counted on to take a population to an efficient outcome of a game in the exceptional case when all of its equilibria happen to be efficient. Perfectly competitive markets are one such case, but markets didn't exist when our species separated itself from the other apes. So how did our unique style of cooperation evolve?

Because relatives share genes, it is easy to explain the evolution of cooperation within the family. For example, any of my genes has half a chance of being present in the body of my sister. If I were genetically programmed to maximize the average number of copies of my genes that are transmitted to the next generation, I would therefore count each of my sister's children as being worth half of one of my own.

This is presumably why some birds help bring up their nephews and nieces when their own chances of raising a family are not very promising.

But human cooperation is more complex. The fierce loyalties that sometimes develop in street gangs or army platoons can perhaps be explained in terms of the collective serving as a surrogate family, but we often manage to cooperate very successfully with total strangers, or with people whom we actively dislike or despise. What is the secret of this seeming contradiction to the doctrine of the selfish gene?

In 1976, Robert Trivers [12] offered "reciprocal altruism" as the solution of the mystery, but David Hume [9] was already on the ball in 1739. As Hume explains:

> I learn to do service to another, without bearing him any real kindness, because I foresee, that he will return my service in expectation of another of the same kind, and in order to maintain the same correspondence of good offices with me and others. And accordingly, after I have serv'd him and he is in possession of the advantage arising from my action, he is induc'd to perform his part, as foreseeing the consequences of his refusal.

**The folk theorem.** Reciprocal altruism cannot work unless people interact repeatedly, without a definite end to their relationship in sight. If the reason I scratch your back today is that I expect you will then scratch my back tomorrow, then our cooperative arrangement will unravel if we know that there will eventually be no tomorrow.[5]

The simplest kind of game in which reciprocity can appear is therefore a *repeated* game with an indefinite time horizon. The simplest of the folk theorems characterizes all the equilibria of such a game in the case when nobody can conceal any information, and everybody always cares about tomorrow nearly as much as they care about today. The important point is that any efficient outcome of the original game on which the players might like to agree approximates an *equilibrium* outcome of the repeated game.

For example, a Pareto-efficient outcome in the one-shot Prisoners' Dilemma is unattainable because the only Nash equilibrium in the game requires that both players defect. But if the Prisoners' Dilemma is repeated indefinitely often, the folk theorem says that cooperation can be sustained as a *self-policing* social contract. In equilibrium, neither player will try to improve their lot by cheating on the cooperative social contract today, because they see that the other player will respond by nullifying today's gain in the future.

The folk theorem improves on David Hume's insight that a social contract is like a masonry arch or a drystone wall by allowing us to examine the details of the equilibrium strategies that sustain it. We can thereby break with the tradition that takes concepts like authority, duty and trust as axiomatic when *explaining* social contracts, and see them rather as words that have evolved to *describe* different social contracts.

[5] This analysis is denied by some behavioral economists, who favor what they call *strong* reciprocity—that people reciprocate because a liking for reciprocation happens to be built into their utility functions (Gintis [7]).

For example, it is fashionable nowadays to attribute our current social woes to a lack of social capital. The implication is that the cure is to inject more social capital into the body politic. But looking around for more social capital is like sending a rookie out for a pint of elbow grease. Social capital isn't a $thing$—it is just a word we use when talking about the properties of an equilibrium that has evolved along with our game of life. Similarly, if we want to know why a citizen obeys an officer of the state, it is not an explanation to say that the citizen respects the authority of the officer. One might as well say that the stones in a masonry arch stay where they are because they do not move.

Just as it is held to be wicked to say that cooperation is possible without people acting selflessly, so the idea that the folk theorem can explain the workings of supposedly difficult ideas like authority, duty, and trust is held to be naive. Let me therefore hasten to explain that game theorists do not think that the ways these notions operate in any given social contract are simple. The detailed workings of real social contracts are complex beyond our capacity to imagine. But traditional moral pundits are too busy making easy things difficult to come anywhere near addressing these fundamental issues. They are right that it would be naive to think that the folk theorem teaches us more than one small secret about human sociality, but this secret, small though it be, is all that is needed to settle much of the dust that these same pundits kick around so furiously.

**Evolution of cooperation.** Games commonly have many equilibria. For example, the Driving Game has three. Such a multiplicity of equilibria in the game of life creates an equilibrium selection problem for evolution to solve if she is to get the players organized into a properly functioning society. In the Driving Game, for example, how is she to replace the inefficient social contract in which drivers randomize by an efficient social contract in which everyone drives on the same side of the road?

The folk theorem tells us that this problem of multiple equilibria was much worse in the repeated games played by our prehuman ancestors. Not only are all the efficient outcomes on which rational players might like to agree available as equilibrium outcomes, but so are large numbers of inefficient equilibria. Why should we expect evolution to succeed in selecting one of the efficient equilibria rather than one of the many inefficient alternatives?

The answer postulates competition among groups. Suppose that many identical small societies are operating one of two social contracts, $a$ and $b$. If $a$ makes each member of a society that operates it fitter than the corresponding member of a society that operates $b$, then here is an argument which says that $a$ will eventually come to predominate.

To say that a citizen is fitter in this context means that the citizen has a larger number of children on average. Societies operating social contract $a$ will therefore grow faster. Assuming societies cope with population growth by splitting off colonies which inherit the social contract of the parent society, we will then eventually observe

large numbers of copies of societies operating social contract $a$ compared with those operating contract $b$.

This retelling of the standard evolutionary story is unusual in two respects. The first is that selection takes place among groups. So why is it not an example of the group selection fallacy? The reason is that a social contract is identified with an $equilibrium$ of the game of life played by each of the competing societies. But selection among equilibria doesn't require that individuals sacrifice anything for the public good, because every individual in every group is already optimizing his or her fitness by acting in accordance with the social contract of his society. The paradigm of the selfish gene is therefore maintained throughout.

The second unusual feature of the story is that the social contract of a parent society is transmitted to its colonies by $cultural$ rather than genetic means. It is hard to convince critics who cannot free themselves of the Lockian blank-slate paradigm of the importance that orthodox evolutionary theorists attach to environmental influences, of which culture is only one. It is a myth that scholars who appeal to evolutionary arguments are "genetic determinists". I know of nobody at all who fits this description. It is uncontroversial that our species somehow learned to use culture as a form of collective unconscious or group mind within which to store the fruits of trial-and-error experimentation from the past, and to incorporate new discoveries made by individuals in the present. Such a cultural resource allows a group to react flexibly in the face of new challenges or opportunities, and hence creates a larger metaphorical cake for the group than would be possible if everybody only knew what they could find out by themselves.

A great deal of our culture is concerned with how we get along with each other; how we split the cake we have jointly created without wasteful conflict. It is this aspect of our social contract with which I am most concerned. Many anthropologists still maintain that only our cultural heritage is relevant to such questions, but it seems obvious to me that human biology must impose constraints on what social contracts can evolve, just as human biology imposes constraints on the deep structure of the languages that can evolve.

David Hume was making a similar point when he observed that the "natural laws" that govern our societies are actually artificial, but are called "natural" because everyone can see that it is natural to the human species that we should have such laws. Similarly, it is natural that a human society should have a language, but its actual language is an artifact of its cultural history. I think that the same is true of social contracts in general. It is natural that a human society should have a social contract, but its actual social contract is an artifact of its cultural history.

If we want to look for $universals$ in the morals of the human species, we therefore have to look beneath the differing cultures of different societies. We must look at the deep structure of human social contracts written into our genes.

Traditionalists are virulently hostile to this suggestion, because they think that nothing is written in our genes but the savagery of nature, red in tooth and claw. I think they are right to the extent that the only social contracts available to us are equilibria in our game of life, but wrong in supposing that such equilibria must

necessarily resemble Thomas Hobbes' state of nature—in which life is "solitary, poor, nasty, brutish and short". On the contrary, the folk theorem tells us that constraining our prehuman ancestors to social contracts that correspond to equilibria in their repeated game of life would have been no bar at all to the evolution of efficient cooperation in their societies.

# 3 Fairness

The folk theorem tells us that many efficient equilibria are available as possible social contracts. To operate successfully, a society needs to single out one of these on which to coordinate. I argue that fairness is evolution's solution to the equilibrium selection problem for the food-sharing aspect of our ancestral game of life. What evidence is there for this conjecture?

All the societies studied by anthropologists that survived into modern times with a pure hunter-gathering economy had similar social contracts with a similar deep structure. This applies across the world—to Kalahari bushmen, Greenland eskimos, Australian aborigines, and Brazilian indians. They tolerate no bosses, and they share on a very egalitarian basis.

Although this global behavior is presumably genetically determined, we are obviously not helpless victims of a genetic predisposition to live in utopian anarchies in which each contributes according to his abilities and receives according to his need. As the economic means of production of a society becomes more complex, its social contract must necessarily adapt if the potential gains from improved technology and the division of labor are to be efficiently exploited. But all the many adaptations that history records are almost certain to have a cultural origin, since the time spans are too short for a biological explanation to be plausible.

Anthropologists started gathering quantitative evidence about food sharing only when pure hunter-gathering societies were on their last legs. So we do not know how important culture was in determining precisely how much was thought to be fair for different people in different societies. But if the general assessment of the way that fairness norms work in modern societies offered in such books as Elster's [6] *Local Justice* or Young's [15] *Equity* is a reliable guide, then there is currently a great deal of cross-cultural variation. So it can't be that the fairness norms we use are determined entirely by our genes.

What counts as fair in different societies can vary as much as the languages spoken. Just as different dialects may be used in different regions of a basically monoglot country, so different views of what counts as fair may operate in different societies, or in the same society at different places or times. Individuals of a particular society even operate different standards when interacting with different groups, or with the same group in a different context—just as schoolkids speak a teenage argot to each other, but communicate with their parents quite comprehensibly.

I have been arguing that what we count as fair depends both on our culture and on our genes. Since cultures vary, any *universal* principles of justice—its deep

structure—must presumably be written into the genes that we all share as members of the same species. If I am right in guessing at the existence of such a deep structure, the next question asks itself. What shape does the deep structure of fairness take?

**The original position.** My book *Just Playing* (Binmore [2]) defends the thesis that the common deep structure of human fairness norms is captured in a stylized form by an idea that John Rawls [11] called the device of the *original position* in his celebrated *Theory of Justice.*

Rawls uses the original position as a hypothetical standpoint from which to make judgments about how a just society would be organized. Members of a society are asked to envisage the social contract to which they would agree *if* their current roles were concealed from them behind a "veil of ignorance". Behind this veil of ignorance, the distribution of advantage in the planned society would seem determined as though by a lottery. Devil take the hindmost then becomes an unattractive principle for those bargaining in the original position, since you yourself might end up with the lottery ticket that assigns you to the rear.

Rawls defends the device of the original position as an operationalization of Immanuel Kant's categorical imperative, but I think this is just window-dressing. The idea certainly hits the spot with most people when they hear it for the first time, but I do not believe this is because they have a natural bent for metaphysics. I think it is because they recognize a principle that matches up with the fairness norms that they actually use every day in solving the equilibrium selection problem in the myriads of small coordination games of which daily life largely consists.

It is important to emphasize that I am not following Rawls here in talking about the grand coordination problems faced by a nation state. Our sense of fairness didn't evolve for use on such a grand scale. Nor am I talking about the artificial and unrealistic principles of justice promoted by self-appointed moral pundits, to which people commonly offer only lip service. I am talking about the real principles that we actually use in solving everyday coordination problems.

The sort of coordination problems I have in mind are those that we commonly solve without thought or discussion—usually so smoothly and effortlessly that we do not even notice that there is a coordination problem to be solved. Who goes through that door first? How long does Adam get to speak before it is Eve's turn? Who moves how much in a narrow corridor when a fat lady burdened with shopping passes a teenage boy with a ring through his nose? Who should take how much of a popular dish of which there isn't enough to go around? Who gives way to whom when cars are maneuvering in heavy traffic? Who gets that parking space? Whose turn is it to wash the dishes tonight? These are picayune problems, but if conflict arose every time they needed to be solved, our societies would fall apart.

Most people are surprised at the suggestion that there might be something problematic about how two people pass each other in the corridor. When interacting with people from our own culture, we commonly solve such coordination problems

so effortlessly that we do not even think of them as problems. Our fairness program then runs well below the level of consciousness, like our internal routines for driving cars or tying shoelaces. As with Molière's Monsieur Jourdain, who was delighted to discover that he had been speaking prose all his life, we are moral in small-scale situations without knowing that we are moral.

Just as we only take note of a thumb when it is sore, we tend to notice moral rules only when attempts are made to apply them in situations for which they are ill-adapted. We are then in the same position as Konrad Lorenz when he observed a totally inexperienced baby jackdaw go through all the motions of taking a bath when placed on a marble-topped table. By triggering such instinctive behavior under pathological circumstances, Lorenz learned a great deal about what is instinctive and what is not when a bird takes a bath. But this vital information is gained only by avoiding the mistake of supposing that bath-taking behavior confers some evolutionary advantage on birds placed on marble-topped tables.

Similarly, one can learn a lot about the mechanics of moral algorithms by triggering them under pathological circumstances—but only if one doesn't make the mistake of supposing that the moral rules are adapted to the coordination problems they fail to solve.[6] However, it is precisely from such sore-thumb situations that I think traditional moralists unconsciously distill their ethical principles. We discuss these and only these situations endlessly, because our failure to coordinate successfully brings them forcefully to our attention.

**Egalitarian or utilitarian?** In arguing that Rawls' original position is built into the deep structure of human fairness norms, I have two tasks. The first task is to offer a plausible account of the evolutionary pressures that might have resulted in such a mechanism being written into our genome. The second task is to explain how this biological mechanism interacts with our cultural heritage to generate a specific choice of equilibrium in some of the games of life we play.

To summarize my approach to the first task would take us to far from the purpose of the current paper. However, I think it necessary to say at least something about the second task, because the orthodox literature on the original position points in two directions that are commonly thought to be diametrically opposed. John Rawls [11] argues that, after the basic rights and liberties of each citizen have been secured, using the device of the original position will lead to an egalitarian distribution of goods and services. John Harsanyi [8]—who independently proposed the device of the original position around the same time as Rawls—argues that its use will lead to a utilitarian distribution.[7]

My naturalistic approach reconciles Harsanyi's and Rawls' conclusions to some extent, but the reconciliation is achieved at the expense of de-Kanting their ideas into a Humean bottle—an activity which both Harsanyi and Rawls viewed with only

[6] A possible mistake that behavioral economists would do well to at least consider.

[7] Harsanyi attributes the idea to William Vickrey. The philosopher, Robert Hare, gives credit for his own version of the idea to the philosopher C. I. Lewis.

cautious sympathy. However, adopting such a new perspective not only allows us to dispense with all metaphysical reasoning, it also allows us to relate the human capacity for empathy to our use of fairness as a coordinating device.

**Empathy and interpersonal comparison.** The original position might be said to be a do-as-you-would-be-done-by principle that takes account of the objection that you shouldn't do unto others as you would have them do unto you, because they may not have the same tastes as yours. For example, Adam may be a keen jogger who likes to be woken before dawn for a ten-mile run through the ice and snow, but Eve is unlikely to respond well if he shakes her awake before the sun has risen on the grounds that this is what he would like her to do for him.

The original position forces Adam to take account of Eve's tastes in such situations by requiring him to consider how it would be if he were to emerge from behind the hypothetical veil of ignorance occupying her role in life. He must therefore have the capacity to empathize with her by putting himself in her position to see things from her point of view. She must simultaneously be able to empathize with him by putting herself in his shoes to see things from his point of view.

Harsanyi [8] saw that to get a grip on what it means for two rational people to empathize with each other is to solve the long-standing puzzle of how interpersonal comparisons of utility can sensibly be made. He thereby created a simple theory in which Adam and Eve's welfare can be measured in a way that makes fairness comparisons possible. He sought to explain the standard of interpersonal comparison that arises in metaphysical terms, but I think that this standard is an artifact of the cultural history of a society.

This claim that our standards of interpersonal comparison of utility are culturally determined is the second pillar of my analogy between fairness and language. Elster [6] and Young [15] document how widely fairness norms can differ in different times and places, but their implicit assumption in describing the differing norms is that the standard of interpersonal comparison is unproblematic, and it is the basic structure of the norm that varies. My alternative explanation for at least some of the variation is that our biology guarantees that the deep structure is always the device of the original position, leaving differences in the observed norms to be explained largely by cultural or contextual variations in the standards of interpersonal comparison.

But how are we to incorporate such considerations into our analysis of how Adam and Eve will bargain in the original position? What relevance do they have for the debate between Harsanyi and Rawls? Should we expect everyday fairness norms to be egalitarian or utilitarian?

Game theory comes to our aid again at this point, because its remit includes the analysis of bargaining. It turns out that much hinges on what one assumes about how the hypothetical deal reached in the original position is enforced. Both Harsanyi and Rawls invent a metaphysical external enforcement agencies whose function is to take up this task. Harsanyi calls his agency "moral commitment". Rawl's calls his agency "natural duty". If the external enforcement agency operates perfectly,

then Harsanyi shows that the final outcome of bargaining in the original position will be a utilitarian outcome in the real world. Rawls evades this conclusion only by the iconoclastic expedient of denying orthodox decision theory.

I have no time for metaphysical inventions. I think Harsanyi's utilitarian conclusion makes good sense when there is a real external enforcement agency—like an all-powerful government whose aim is to enforce the laws that the citizens would choose for themselves under fair conditions. Some modern governments arguably seek to approach this ideal, but there was nothing evenly remotely like such an external enforcement agency when our prehuman ancestors evolved a sense of fair play. So what will be the outcome of bargaining in the original position when no external enforcement is assumed? Everything will then need to be in equilibrium, including the process by means of which a society moves from an inefficient equilibrium to a Pareto-improving fair equilibrium.[8]

My answer justifies Rawls' basic intuition about how real fairness norms work. His formal analysis is faulty, but the result in the real world of rational bargaining in the original position will nevertheless be egalitarian—not in exactly the sense he proposed, but in the sense made precise by what game theorists call the egalitarian (or proportional) bargaining solution (Binmore [2]). Aristotle [1] made the essential point long ago when he said:

"What is fair . . . is what is proportional"

A small group of social psychologists who call their subject "modern equity theory" defend the same conclusion empirically using experimental data in which subjects are asked what they regard as fair in various circumstances (Wagstaff [13]).

# 4 Conclusion

The battle to persuade the economics profession that institutions matter was won many years ago. It is no longer necessary—if it ever was—to explain why perfectly competitive markets are not the answer to all economic problems. We can therefore get down to the problem of analyzing how institutions work, both those that already exist and those that exist only in the minds of theorists with utopian aspirations.

The view I have tried to defend in this paper applies when one steps back from the detailed analysis of how particular institutions operate to examine why they evolved and how they survive. It conceptualizes an institution as a solution to an equilibrium selection problem in games that have many equilibria, thus potentially making the methodology of game theory available to institutional economists on a wider stage than previously.

Many problems in game theory remain unresolved. A particularly important unanswered question for institutional economics is the extent to which the folk theorem of repeated game theory survives when the strategic choices of the players

[8] One might say that one thereby takes Rawls' concerns about the "strains of commitment" to their logical extreme.

are only imperfectly monitored by the other players. In cases like this, there is much scope for cross-fertilization between institutional economics and game theory—just as I found scope for cross-fertilization between anthropology and game theory in my attempt to come up with a naturalistic explanation of the human sense of fair play.

It is true that such a melding of the disciplines of institutional economics and game theory would require laying less stress on certain conventional ways of thinking that are usually thought to typify the New Institutional Economics, but most disciplines—including game theory—could benefit from a little creative destruction.

# References


[1] Aristotle. Politics. In J. Barnes, editor, *The Complete Works of Aristotle, Volume II*. Princeton University Press, Princeton, 1984.

[2] K. Binmore. *Just Playing: Game Theory and the Social Contract II*. MIT Press, Cambridge, MA, 1998.

[3] K. Binmore. *Natural Justice*. Oxford University Press, New York, 2005.

[4] K. Binmore. The origins of fair play. *Proceedings of the British Academy*, 151:151–193, 2007.

[5] K. Binmore and A. Shaked. Experimental economics: Where next? See http://www.wiwi.uni-bonn.de/shaked/BS-FS-appendix/, 2009.

[6] J. Elster. *Local Justice: How Institutions Allocate Scarce Goods and Necessary Burdens*. Russell Sage Foundation, New York, 1992.

[7] H. Gintis. *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton University Press, Princeton, 2009. (Forthcoming).

[8] J. Harsanyi. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press, Cambridge, 1977.

[9] D. Hume. *A Treatise of Human Nature (Second Edition)*. Clarendon Press, Oxford, 1978. (Edited by L. A. Selby-Bigge. Revised by P. Nidditch. First published 1739).

[10] D. North. *Understanding the Process of Institutional Change*. Princeton University Press, Princeton, 2005.

[11] J. Rawls. *A Theory of Justice*. Oxford University Press, Oxford, 1972.

[12] R. Trivers. The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46:35–56, 1971.

[13] G. Wagstaff. *An Integrated Psychological and Philosophical Approach to Justice*. Edwin Mellen Press, Lampeter, Wales, 2001.

[14] O. Williamson. The new institutional economics: Taking stock, looking ahead. *Journal of Economic Literature*, 38:595–613, 2000.

[15] P. Young. *Equity*. Princeton University Press, Princeton, 1994.