

Do Conventions Need To Be Common Knowledge?

Ken Binmore
Economics Department
University College London
Gower Street
London WC1E 6BT, UK

Do Conventions Need To Be Common Knowledge?

by Ken Binmore

1 Conventions

Two men who pull the oars of a boat, do it by an agreement or convention, although they have never given promises to each other. Nor is the rule concerning the stability of possessions the less derived from human conventions, that it arises gradually, and acquires force by a slow progression, and by our repeated experience of the inconveniences of transgressing it. On the contrary, this experience assures us still more, that the sense of interest has become common to all our fellows, and gives us confidence of the future regularity of their conduct; and it is only on the expectation of this that our moderation and abstinence are founded. In like manner are languages gradually established by human conventions without any promise. In like manner do gold and silver become the common measures of exchange, and are esteemed sufficient payment for what is of a hundred times their value.

David Hume's [11, III,II,ii] wisdom in this famous passage was not appreciated by his contemporaries. It was only with the advent of game theory some 200 years later that less able folk were provided with a crutch that allowed them to walk where he had run. Nowadays, we are able to follow Thomas Schelling [22] in regarding a convention as a social device whose function is to coordinate our actions on one particular equilibrium when the game that life calls on us to play has multiple equilibria.

When two men row a boat, it isn't in equilibrium for one man to row more strongly than the other, because the boat will then go round in a circle. If this happens, each man will prefer to change his rowing rhythm to match that of his partner. If they succeed in doing so, they will have reached an equilibrium. But the rowing game admits many possible equilibria. The equilibrium they actually adopt is a *convention* for their game. It may be unique to the minisociety consisting of just the two rowers, or it may be a convention shared by a whole community of rowers. Either way, it is a cultural artifact that might have been different without contravening any principle of individual rationality.

Metaphysics? Most people have no difficulty in accepting the conventional nature of language or money, but draw the line when the philosophers like Hume suggest that the same is true in more sensitive subjects like ethics or religion. In his *Conventions*, David Lewis [13] boldly extends the argument even to epistemology, essentially arguing that the boundaries of what we call conventional wisdom need to be set far wider than convention currently allows.

My own view is that caution is necessary when arguing that culture is free to make anything whatever into a convention. For example, Chomsky has shown that all human languages have a common deep structure which is presumably written into our genes. I argue elsewhere that the same may be true of human fairness norms (Binmore [4]). However, I am completely sold on the idea that conventions run much deeper than is generally accepted, and that we shall never understand how human societies work as long as we continue to confuse pieces of conventional wisdom that are products of our biological and cultural history with metaphysical principles carved into the fabric of the universe.

Evolutionary and educative game theory. Game theory splits into two branches that reflect the same philosophical divide. I call the two branches evolutionary and educative game theory (Binmore [2]). The players in evolutionary game theory need not be thinking creatures at all. In some of the more successful applications, they are plants or insects. Insofar as game theory is able to predict their behavior, it is because some process of trial-and-error adjustment kept moving the ecology of which they are a part until it settled down into an equilibrium of their underlying game of life. Similarly, we do not imagine that the men who row Hume's boat will think at all deeply about how they should row. We take for granted that they will unconsciously adjust their rhythm until the boat is moving smoothly through the water in the direction they wish to go.

The tradition in game theory inherited from Von Neumann and Morgenstern [26] that I call educative could equally well be called rationalistic. Axioms are proposed that supposedly govern the behavior of ideally rational players. The behavior of different agents is linked by hypothesizing that they know relevant things about each other and the game they are playing. Their behavior in the game is then deduced from these assumptions.

The first jewel in the crown of educative game theory was Von Neumann's minimax theorem for two-person, zero-sum games, which says that it is optimal for both players to choose a strategy on the apparently paranoid assumption that their opponent will guess their choice and act to minimize their payoff.

Many strategies commonly satisfy this requirement, but this is not a problem for Von Neumann, because it doesn't matter how one solves the equilibrium selection problem for two-person, zero-sum games.¹ It is perhaps because there was no equilibrium selection problem in the original domain of game theory that it took a relative outsider like Schelling [22] to make it clear that conventions are inescapable in the general case.

In this paper, I plan to argue—contra David Lewis [13]—that the study of conventions is more fruitfully pursued from the foundations on which evolutive game theory is based rather than the much more demanding foundations of eductive game theory. It is true that Hume tells us (immediately before the passage quoted at the head of this section) that a human convention should be understood as “a general sense of common interest” from which “a suitable resolution and behaviour” follows when “it is mutually expressed and is known to us both”. However, I shall argue that accepting Lewis’s attempt to formalize Hume’s remarks in terms of what should and should not be construed as common knowledge would make it very hard indeed for new conventions to get established. Rather than invent a new term for the coordination devices that do succeed in colonizing a society, I therefore argue for a more relaxed attitude to what should count as a convention.

2 Game Theory

Figure 1 shows four payoff tables for some canonical toy games. I call the two players Alice and Bob. In each game, Alice has two strategies represented by the rows of the payoff table. Bob also has two strategies represented by its columns. The four cells of the payoff table correspond to the possible outcomes of the game. Each cell contains two numbers, one for Alice and one for Bob. The number in the southwest corner is Alice’s payoff for the corresponding outcome of the game. The number in the northeast corner is Bob’s payoff.

In Matching Pennies, each player shows a coin. Alice wins if they are both the same, and Bob if they differ. The payoffs in each cell of Matching Pennies add up to zero. One can always fix things to make this true in such a game of pure conflict. For this reason, games of pure conflict are said to be zero sum.

We play the Driving Game every time we get in our cars to drive to work in the morning. The payoffs in each cell of the Driving Game are equal. One can

¹All equilibria in two-person, zero-sum games are interchangeable and payoff-equivalent.

always fix things to make this true in such a game of pure coordination.

The Prisoners' Dilemma and the Stag Hunt Game lie between the two extremes of pure conflict and pure coordination, with the *dove* strategy representing cooperation and the *hawk* strategy representing defection. In spite of Brian Skyrms' book [24] *The Stag Hunt*, the Stag Hunt Game seems to be less widely known than the Prisoners' Dilemma. It derives from a story of Jean-Jacques Rousseau [20] in which Alice and Bob agree to cooperate in hunting a stag. When they separate to put their plan into action, each may be tempted to abandon the joint enterprise by the prospect of bagging a hare for themselves.

| | <i>heads</i> | <i>tails</i> |
|--------------|--------------|--------------|
| <i>heads</i> | 1* -1 | -1 1* |
| <i>tails</i> | -1 1* | 1* -1 |

Matching Pennies

| | <i>dove</i> | <i>hawk</i> |
|-------------|-------------|-------------|
| <i>dove</i> | 5* 5* | 0 4 |
| <i>hawk</i> | 4 0 | 2* 2* |

Stag Hunt Game

| | <i>left</i> | <i>right</i> |
|--------------|-------------|--------------|
| <i>left</i> | 1* 1* | 0 0 |
| <i>right</i> | 0 0 | 1* 1* |

Driving Game

| | <i>dove</i> | <i>hawk</i> |
|-------------|-------------|-------------|
| <i>dove</i> | 3 3 | 0 4* |
| <i>hawk</i> | 4* 0 | 2* 2* |

Prisoners' Dilemma

Figure 1: Some toy games.

Nash equilibrium. Each player is assumed to seek to maximize his or her expected payoff in a game. This would be easy if a player knew what strategy the other were going to choose. For example, if Alice knew that Bob were going to choose *left* in the Driving Game, she would maximize her payoff by choosing *left* as well. That is to say, *left* is Alice's best reply to Bob's choice of *left*, a fact indicated in Figure 1 by starring Alice's payoff in the cell that results if both players choose *left*.

A cell in which *both* payoffs are starred corresponds to a Nash equilibrium, because each player is then simultaneously making a best reply to the strategy

choice of the other (Nash [18]). Sometimes players have more than one best reply, but if both players make best replies to each other that are strictly better than all their alternatives, a Nash equilibrium is said to be strict.

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 a Nash equilibrium. If it weren't, 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 Nash equilibria matter is equally important. If the payoffs in a game correspond to how fit the players are, then evolutionary processes—either cultural or biological—that favor strategies that currently generate a higher payoff at the expense of those that generate a lower payoff will stop working when we get to an equilibrium, because all the surviving strategies will then be as fit as it is possible to be in the circumstances. Only Nash equilibria can therefore be evolutionarily stable.²

Much of the power of game theory as a conceptual tool derives from the possibility of moving back and forward between these eductive and evolutive interpretations of an equilibrium.

The security dilemma. Both payoffs are starred in two cells of the payoff table of the Driving Game, and so both these cells correspond to Nash equilibria. It is an equilibrium if everyone drives on the left. It is also an equilibrium if everyone drives on the right. The players get the same payoff at each of these equilibria and so they don't care whether they both drive on the left or they both drive on the right. Their only concern is that they both coordinate on the same equilibrium. However, the same isn't true of the Stag Hunt Game.

Like the Driving Game, the Stag Hunt Game has two cells in which both payoffs are starred. Both of these cells correspond to Nash equilibria, but now both players prefer the cooperation equilibrium in which both play *dove* to the defection equilibrium in which both play *hawk*. If they live in a society in which it is conventional to play the cooperation equilibrium then all is well, but suppose defection is the established convention.

What can rational players do to persuade each other before playing the game

²John Maynard Smith [14] defines an evolutionarily stable strategy as a best reply to itself that is a better reply to any alternative best reply than the alternative best reply is to itself, but biologists don't seem to worry much about the small print involving alternative best replies.

that their minisociety should shift to the cooperation convention? Experts in international relations study the Stag Hunt Game under the name of the Security Dilemma because it is the simplest case where the answer to this question is problematic. It is important to realize that the problem isn't just a question of one player pointing out the advantages of shifting to a new convention. Alice may tell Bob that she plans to play *dove* on the assumption that he will be convinced by her arguments in favor of the new convention, but will Bob believe her?

Whatever Alice is planning to play, it is in her interests to persuade Bob to play *dove*. If she succeeds, she will get 5 rather than 0 when playing *dove*, and 4 rather than 2 when playing *hawk*. Rationality alone therefore doesn't allow Bob to deduce anything about her plan of action from what she says, because she is going to say the same thing no matter what her real plan may be! Alice may actually think that Bob is unlikely to be persuaded to switch from *hawk* and hence be planning to play *hawk* herself, yet still try to persuade him to play *dove*.

The point of this Machiavellian story is that attributing rationality to the players isn't enough to resolve the equilibrium selection problem—even in a case that seems as transparently straightforward as the Stag Hunt Game. If we see Alice and Bob playing *hawk* in the Stag Hunt Game, we may regret their failure to coordinate on playing *dove*, but we can't accuse either player of being irrational, because neither player can do any better given the behavior of their opponent.

It doesn't help to point out the obvious fact that Alice and Bob would both be better off if both had more faith in each other's honesty. Nobody says that it is irrational for people to trust each other. Game theorists only say that it isn't rational to trust people without a good reason—that trust can't be taken on trust. Or, to quote David Hume [11] again, "Surely I am not bound to keep my word because I have given my word to keep it."

Of course, there is usually a lot more going on in the real world than in the highly idealized microcosm of a formal game. For example, Sweden switched from driving on the left to driving on the right in the early hours of September 1st, 1967. But who thinks that the notoriously misanthropic Iks would have responded similarly to a call from the Ugandan government to shift to a more cooperative equilibrium of their tribal game of life? (Turnbull [25].)

Categorical imperative? The Prisoners' Dilemma is mentioned by way of counterpoint to the Stag Hunt Game. It has the same payoffs, except that mutual cooperation has been made less attractive by reducing the payoffs the players receive at the cooperation outcome from 5 to 3. The result is that *hawk* now

strictly dominates *dove*, which means that *hawk* is a strict best reply whatever strategy the other player may choose. In particular, the Prisoners' Dilemma has only one Nash equilibrium, in which both players choose *hawk*.

Immanuel Kant's categorical imperative would seem to contradict the claim that only the play of *hawk* is rational in the Prisoners' Dilemma. In fact, Kant is only one of many scholars who have argued that play can be rational without being in equilibrium (Binmore [3, Chapter 3]). It may therefore be worthwhile to clarify the Humean sense in which game theorists understand rational play.

So as not to beg any questions, we begin by asking where the payoff table that represents the players' preferences in the Prisoners' Dilemma comes from. The game theory answer is that we discover the players' preferences by observing the choices they make (or would make) when solving one-person decision problems.

Writing a larger payoff for Alice in the bottom-left cell of the payoff table of the Prisoners' Dilemma than in the top-left cell therefore means that Alice would choose *hawk* in the one-person decision problem that she would face if she knew in advance that Bob had chosen *dove*. Similarly, writing a larger payoff in the bottom-right cell means that Alice would choose *hawk* when faced with the one-person decision problem in which she knew in advance that Bob had chosen *hawk*. The very definition of the game therefore says that *hawk* is Alice's best reply when she knows that Bob's choice is *dove*, and also when she knows his choice is *hawk*. So Alice doesn't need to know anything about Bob's actual choice to know her best reply to it. It is rational for her to play *hawk* whatever strategy he is planning to choose.

Mixed strategies. The payoff table of Matching Pennies has no cell with both payoffs starred. It follows that the game has no Nash equilibrium in pure strategies. But the players aren't restricted to playing *left* or *right*. They can also mix between these strategies by randomizing their choice.

Nash [18] proved that all finite games have at least one Nash equilibrium when such mixed strategies are allowed. Matching Pennies has a unique mixed equilibrium that requires each player to choose *left* or *right* with equal probability. Since Matching Pennies is a zero-sum game, this strategy is the same as Von Neumann's paranoid strategy. If Alice plays *left* or *right* with equal probability, she is sure to win half the time on average, whatever Bob may do.³ Since the same is true of Bob, both players will be making a best reply to the

³When outcomes other than just winning or losing can arise, it is necessary to interpret the payoffs as Von Neumann and Morgenstern [26] utilities.

(mixed) strategy choice of their opponent if they both play *left* or *right* with equal probability.

Both the Driving Game and the Stag Hunt Game also have a mixed-strategy equilibrium as well as their two pure-strategy equilibria. Such a multiplicity of equilibria is typical of more realistic games. Matching Pennies and the Prisoners' Dilemma are unusual in not posing an equilibrium selection problem.

Lewis on game theory. David Lewis's [13] game theory is not very orthodox, and so some clarification may be helpful.

Lewis [13, p.8] does not mention John Nash, but what he calls an equilibrium is a Nash equilibrium in pure strategies. He doesn't consider mixed Nash equilibria. What he calls a proper equilibrium is not a proper equilibrium in the sense of Myerson [16], but what is normally called a strict Nash equilibrium. Lewis [13, p.17] offers only one formal proof: that a game of pure coordination with a unique Nash equilibrium must have a dominated strategy. Figure 2(a) shows this to be false with the standard definition of a dominated strategy, but Lewis uses a nonstandard definition.⁴

| | | | |
|---|---|---|---|
| | 4 | 3 | 0 |
| 4 | 4 | 3 | 0 |
| 0 | 0 | 2 | 1 |
| 3 | 3 | 0 | 2 |

(a)

| | | | |
|----|----|---|----|
| | -1 | 0 | 1 |
| 1 | 1 | 0 | -1 |
| 0 | 0 | 0 | 0 |
| -1 | 1 | 0 | -1 |

(b)

Figure 2: A game of pure coordination and a game of pure conflict.

Lewis's [13, p.14] definition of a coordination equilibrium is also eccentric. He notes that an equilibrium is a combination of strategies in which no one would have been better off if he alone had acted otherwise. He then says that "a coordination equilibrium is a combination of strategies in which no one would have been better off if *any one* agent alone acted otherwise, either himself or

⁴He says that a pure strategy is strictly dominated if it is never a best reply to any strategy combination available to the other players. With this weak definition, it is false that a strictly dominated strategy is never used with positive probability in equilibrium.

someone else.” Rather than use this definition, I shall not speak of coordination equilibria at all.

Lewis’s definition makes the unique equilibrium of Matching Pennies into a coordination equilibrium, and it seems perverse to speak of coordination in a game in which each player’s aim is to *prevent* the opponent’s attempt to coordinate their strategies. If it is objected that the equilibrium is mixed in Matching Pennies, one can make the same point with the game of Figure 2(b), in which a pure strategy has been introduced that has the same effect as playing *left* or *right* with equal probabilities. (Lewis [13, p.15] uses the same example without its first row and column.)

3 Nash Demand Game.

Schelling [22] conducted some instructive experiments in the 1950s on how people manage to solve various games of pure and impure coordination. In his best known experiment, the subjects were asked what two people should do if they had agreed to meet up in New York tomorrow without specifying a place and time in advance. The standard answer was that they should go to Grand Central Station at noon. When people commonly agree on such a resolution of a coordination problem, Schelling says that the consensus they report constitutes a focal point. If it were necessary to distinguish between a focal point and a convention, perhaps the criterion would be that a focal point is a convention that the players aren’t aware that they are likely to share in advance of playing a coordination game.

A politically incorrect version of Schelling’s meeting problem is traditionally called the Battle of the Sexes. Adam and Eve are a pair of honeymooners who get separated in a big city after failing to agree at breakfast on whether to meet up at the ballet or a boxing match. Alice prefers the former and Bob the latter. The Nash Demand Game [17] can be regarded as a more elaborate version of this game in which partial coordination is also possible.

However, the Nash Demand Game is more commonly interpreted as a primitive bargaining model in which the feasible payoff pairs lie in a set X like that shown in Figure 3. Alice and Bob each simultaneously demand a payoff. If the pair of payoffs demanded is in the feasible set, both players receive their demands. If not, both players receive the disagreement payoff of zero.

The game poses the equilibrium selection problem in an acute form, because every efficient outcome that assigns both players no less than their disagreement payoffs is a Nash equilibrium of the game. For this reason, the game has become

a standard testbed for trying out equilibrium selection ideas (Skyrms [23]). In Section 6, we explore the fairness conventions that people may come to regard as appropriate in this game.

Nash himself proposed dealing with the equilibrium selection problem strategically by studying a smoothed version of the Nash Demand Game in which the players aren't certain where the boundary of the feasible set starts and stops. As one moves out along a curve from the disagreement point, the probability that the current payoff pair is feasible declines smoothly from one to zero in the vicinity of the boundary. All the Nash equilibria of the unsmoothed game are still approximate equilibria of the new game, but the *exact* Nash equilibria of the smoothed game all lie near a payoff pair N called the Nash bargaining solution that Nash [17] famously characterized axiomatically.

4 Common Knowledge

In a discussion of how particular equilibria in games of pure coordination become focal, Schelling [22, p.109] proposes a mind experiment in which the players are envisaged as being connected to machines that register the focus of their attention: "Each can see the meter on his own machine, each can see the meter on the other's machine, and each is aware that both are aware that both can see both meters."

Lewis [13, p.58] took this line of thought further by insisting that a convention can only operate in an informational environment with such a character. After reminding us that his conception of the nature of a convention requires a regularity in behavior, a system of mutual expectations, and a system of preferences, he then insists that these properties must be common knowledge in the population in which the convention is established. His formal expression of the latter requirement is reproduced below:

A regularity R in the behavior of members of a population P when they are agents in a recurring situation S is a *convention* if and only if it is true that, and it is common knowledge in P that, in any instance of S among members of P ,

- (1) everyone conforms to R ;
- (2) everyone expects everyone else to conform to R ;
- (3) everyone prefers to conform to R on condition that the others do, since S is a coordination problem and uniform conformity to R is a coordination equilibrium in S .

Lewis continues by drawing attention to the fact that these requirements set up an infinite chain of expectations in accordance with the current understanding in game theory that something is common knowledge if everybody knows it, everybody knows that everybody knows it, everybody knows that everybody knows that everybody knows it; and so on. Bob Aumann's [1] later definition of common knowledge in terms of the players' knowledge partitions allowed this insight to be put onto a formal basis, but space doesn't permit a discussion of the ingenious manner in which Aumann avoids following Lewis into a tangle of infinite regressions. Instead, I plan to say only enough to make it clear that the propositions of the theory have the status of theorems (Binmore [6, Chapter 12]).

Modeling knowledge. We can specify what Alice knows with the help of a knowledge operator \mathcal{K} . Events are modeled as subsets of a finite set Ω of all states of the world. For each event E , $\mathcal{K}E$ is the set of states of in which Pandora knows that E has occurred. That is to say, $\mathcal{K}E$ is the event that Alice knows E .

In the small world created when a game is specified, the knowledge operator \mathcal{K} is assumed to satisfy the requirements of the modal logic S-5. These are listed below, along with the equivalent requirements for the possibility operator \mathcal{P} defined by $\mathcal{P}E = \sim\mathcal{K}\sim E$, where $\sim F$ is the complement of the set F .

- | | | | |
|------|--|------|--|
| (K0) | $\mathcal{K}\Omega = \Omega$ | (P0) | $\mathcal{P}\emptyset = \emptyset$ |
| (K1) | $\mathcal{K}(E \cap F) = \mathcal{K}E \cap \mathcal{K}F$ | (P1) | $\mathcal{P}(E \cup F) = \mathcal{P}E \cup \mathcal{P}F$ |
| (K2) | $\mathcal{K}E \subseteq E$ | (P2) | $\mathcal{P}E \supseteq E$ |
| (K3) | $\mathcal{K}E \subseteq \mathcal{K}^2E$ | (P3) | $\mathcal{P}E \supseteq \mathcal{P}^2E$ |
| (K4) | $\mathcal{P}E \subseteq \mathcal{K}\mathcal{P}E$ | (P4) | $\mathcal{K}E \supseteq \mathcal{P}\mathcal{K}E$ |

Game theorists who are surprised that they believe all these propositions may be comforted at the news that they are equivalent to assuming that the collection of all possibility sets $\mathcal{P}\{\omega\}$ partitions Ω , just like the information sets in a game of imperfect information.

I say that something that cannot be true without Alice knowing it is a truism for her. So T is a truism if and only if $T \subseteq \mathcal{K}T$. By (K2), we then have $T = \mathcal{K}T$. If we regard a truism as capturing the essence of what happens when making a direct observation, it can be argued that all knowledge derives from truisms.

This observation is reflected in the following trivial theorem:⁵

Alice knows that E has occurred if and only if a truism T that implies E has occurred.

Public events. The crudest way to define the common knowledge operator $(\text{everybody knows})^\infty$ is to take $(\text{everybody knows})^\infty E$ to be the intersection of all the sets $(\text{everybody knows})^N E$.

The common knowledge operator inherits all the properties (K0)–(K4) of an individual knowledge operator. In particular, an event can be common knowledge if and only if it is implied by a common truism—an event that cannot occur without its becoming common knowledge. However, it turns out that a common truism is the same thing as a public event, which has a simpler characterization. A subset E of Ω is a public event if and only if it cannot occur without everybody knowing it has occurred, so that

$$E = (\text{everybody knows})E.$$

This observation returns us to Schelling’s mind experiment in which each player is aware that both are aware that both can see both meters. With remarkable prescience, he sets up the conditions of his mind experiment so that the focalizing behavior of the players is a public event. As Lewis then observes, the outcome is that the resulting focal point or convention will be common knowledge between the players.

I don’t want to downplay the importance of public events to the maintenance of human social systems. The significance we attach to making eye contact is enough in itself to show that it matters that some events are public. When Alice refuses to make eye contact with Bob, she is refusing to make it common knowledge between them that they recognize each other as persons. Presumably, this is why we are careful not to make eye contact with beggars when we plan to disregard their need.

However, it seems to me that the main consequence of modern advances in our understanding of knowledge operators for the theory of conventions is

⁵If the true state ω lies in a truism T with $T \subseteq KE$, we first show that Alice knows that E has occurred. But if $\omega \in T \subseteq KE$, then $\omega \in KE$, whether or not T is a truism. We next show that if Alice knows that E has occurred, then a truism T has occurred with $T \subseteq E$. This is easy, because we can just take $T = KE$. The event T is a truism, because (K3) says that $T \subset KT$. The truism T must have occurred because to say that Pandora knows that E has occurred means that the true state $\omega \in KE = T$.

to bring forcibly to our attention how difficult it is for something to become common knowledge. How often do we have the opportunity to observe each other observing something? For large numbers of people, I guess the answer is *never*. So how can a language have become a convention if a convention needs to be common knowledge in a society? How can it have become conventional for gold to be valuable? How can it even have become conventional to drive on the right? And if we do not know the answer to such questions as these, how are we ever to find our way to an equilibrium of the game of life we play on this planet in which we get global warming and the like under control?

This is not to deny that we behave *as if* conventions are common knowledge when we use them, but Lewis's definition would seem to make it impossible for conventions to get started except in very restricted circumstances.

Lewis [13, p.78] relaxes his definition of a convention to the requirement that it only need be common knowledge that a convention is honored by some fraction of a population, but I do not see that this helps with the problem identified above. It is true that Bob need only believe that his opponent will play *dove* more than two thirds of the time in our version of the Stag Hunt Game for it to be optimal for him to play *dove* himself, but Lewis still requires this belief to be common knowledge. A further modification in which the requirement of common knowledge is replaced by Monderer and Samet's [15] notion of common p -belief⁶ would be a more useful response to the problem I am raising, but one would still be left with essentially the same difficulty.

5 Byzantine Generals

In computer science, the difficulties that arise when two people seek to upgrade a piece of knowledge held by one into a piece of knowledge held in common are illustrated by the "coordinated attack problem" Halpern *et al* [9]).

Two Byzantine generals occupy adjacent hills with the enemy in the valley between. If both generals attack together, victory is certain, but if only one general attacks, he will suffer badly. The first general therefore sends a messenger to the second general proposing an attack. Since there is a small probability that any messenger will be lost while passing through the enemy lines, the second general sends a messenger back to the first general confirming the plan to at-

⁶Wherever something is asserted to be known in the standard theory, say instead that it is believed with probability at least p .

tack. But when this messenger arrives, the second general doesn't know that the first general knows that the second general received the first general's message proposing an attack. The first general therefore needs to send another messenger confirming the arrival of the second general's messenger. But when this messenger arrives, the first general doesn't know that the second general knows that the first general knows that the second general received the first general's message. The fact that an attack has been proposed is therefore not common knowledge because, for an event E to be common knowledge, *all* statements of the form (everybody knows that) ^{n} E must be true. Further messengers may be shuttled back and forward until one of them is picked off by the enemy, but no matter how many confirmations each general may receive before this happens, it *never* becomes common knowledge that an attack has been proposed.

This looks like a major problem for the coherence of the distributed systems studied in computer science, because two different smart agents will necessarily have different information as a consequence of their differing experience. How can they act together in a joint enterprise if they cannot succeed in sharing their knowledge adequately?

A clue to the fact that a wrong question is possibly being asked here is to be found in the coordinating behavior of ordinary people. When Alice texts a suggestion to Bob that they meet at noon in the coffee shop and Bob texts the reply OK, this is usually enough to ensure that Alice and Bob will meet up successfully. But their agreement isn't common knowledge between them because Bob didn't get his confirmation confirmed. This commonplace observation suggests that it may be worth reconsidering the arguments which suggest that coordinated action must be based on common knowledge.

It isn't true—as sometimes claimed—that there must be common knowledge of the game and of the players' rationality in order to justify the play of a Nash equilibrium. For two optimizing players to operate a Nash equilibrium, it is obviously sufficient if each knows the strategy that the other plans to play. In the Prisoners' Dilemma, even this much knowledge is superfluous. However, Rubinstein's Email Game would seem to show that such a removal of the focus of the discussion from what the players know to the actions they need to take to implement a Nash equilibrium doesn't eliminate the problem.

The Email Game. Independently of the computer science literature, Rubinstein [21] formulated a version of the coordinated action problem in terms of his Electronic Mail Game. Instead of two Byzantine generals, we have Alice and Bob

communicating by email. They have an opportunity from which they can both profit only if they coordinate on exploiting it. Only Bob knows of the opportunity, and hence must communicate with Alice if the opportunity is to be seized. Bob can send a message to Alice, but there is some probability that her message won't arrive. If the message arrives, Alice sends an acknowledgement which again may fail to arrive. Bob acknowledges the acknowledgement, and so on. The question is whether Alice and Bob will be able to exploit their opportunity.

The Email Game is a formal version of this problem in which Alice and Bob must independently choose between `DOVE` and `HAWK` (where the use of capitals is significant). Their payoffs are then determined by whether Chance makes `DOVE` correspond to *dove*, and `HAWK` to *hawk* in the Stag Hunt Game, or whether she reverses these correspondences. It is common knowledge that Chance chooses the first possibility two thirds of the time.

Only Bob learns what decision Chance has made. On the understanding that the default action is `DOVE`, a message goes to Alice that says "Play `HAWK`" whenever Bob learns that *dove* corresponds to `HAWK`. Alice's machine confirms receipt of the message by bouncing it back to Bob's machine. Bob's machine confirms that the confirmation has been received, by bouncing the message back again. And so on.

The (everybody knows)^N operator applies with ever higher values of N as confirmation after confirmation is received. So if the players could wait until infinity before acting, Chance's choice would become common knowledge. However, the Email Game is realistic to the extent that the probability of any given message failing to arrive is some very small $\epsilon > 0$. The probability of Chance's choice becoming common knowledge is therefore zero. But we can still ask whether coordinated action is possible for Alice and Bob. Is there a Nash equilibrium in which they do better than always playing their default action of `DOVE`?

The possible states of the world are the number of messages that get sent. Neither player knows the actual state of the world. For example, if the state of the world is 2 (so that the third message went astray), then Bob thinks it also possible that the second message (sent by Alice's machine) wasn't sent because the first message (sent by Bob's machine) didn't arrive. Bob's possibility set is therefore $P_B\{2\} = \{1, 2\}$. Similarly, Alice's possibility set is $P_A\{2\} = \{2, 3\}$ when the state of the world is 2.

In the Email Game, a pure strategy specifies an action (either `DOVE` or `HAWK`) for each of a player's possibility sets. Rubinstein showed that the only Nash equilibrium consistent with Bob's choosing `DOVE` when no message is sent

requires both players to choose DOVE at all their possibility sets.⁷ No convention that allows Alice and Bob always to coordinate on the cooperation equilibrium in the Stag Hunt Game is therefore available.

Byzantium rescued! Rubinstein's [21] widely quoted result on the Email Game seems to support Lewis's intuition that common knowledge is necessary for a convention to be operational. No matter how many times we succeed in iterating the (everybody knows) operator, we get no nearer to implementing a fully cooperative convention.⁸

However, one should never put too much weight on a single formal model. It turns out that Rubinstein's conclusion depends on the fact that Alice and Bob's machines automatically bounce back a confirmation when they receive a message. This seems an innocent simplification, but if we allow Alice and Bob the freedom to choose whether or not to send back a confirmation, then the results of the model are turned upside down (Binmore and Samuelson [7]). Instead of full cooperation being unavailable as a Nash equilibrium no matter how many messages are sent, we find that there is a plethora of Nash equilibria that support full cooperation. Whatever positive number of messages is specified in advance, there is a Nash equilibrium in which both players use HAWK after that number of messages have been sent and received.⁹ It therefore turns out that common knowledge is irrelevant to the operation of the convention.

In the most pleasant Nash equilibrium of the modified Email game, both players play HAWK whenever Bob proposes doing so and Alice says OK—as when friends agree to meet in a coffee shop. But there are other Nash equilibria in which the players settle on HAWK only after a long exchange of confirmations

⁷If Alice plays the default action DOVE at $\{0, 1\}$, then it is optimal for Bob to play DOVE at $\{1, 2\}$. On reaching this possibility set, Bob believes it more likely that the state of the world is 1 rather than 2, because the second message can only go astray if the first message is received. Can it then be optimal for him to play HAWK? The most favorable case is when each state is equally likely, and Alice is planning to play DOVE at $\{2, 3\}$. Bob might as well then be playing against someone playing each strategy in the ordinary Stag Hunt Game with equal probability, so his optimal reply is *hawk*, which he knows corresponds to DOVE at $\{1, 2\}$. Similarly, Bob playing DOVE at $\{1, 2\}$ implies that Alice plays DOVE at $\{2, 3\}$. And so on.

⁸Mathematicians would say that there is a discontinuity at infinity. That is to say, when we take the limit as $N \rightarrow \infty$, we don't get the same result as when we set $N = \infty$. (Monderer and Samet [15] argues that we should be taking the limit as $p \rightarrow 1$ of the common p -belief operator.

⁹One can restrict the number of Nash equilibria by imposing costs of sending and receiving messages, but this does not affect the basic result.

of confirmations. Hosts of polite dinner parties suffer from such equilibria when their guests start moving infinitely slowly towards the door at the end of the evening, stopping every inch or so in order that the host and the guest can assure each other that departing at this time is socially acceptable to both sides. One might hope that social evolution would eventually eliminate such long goodbyes, but the prognosis isn't good. Only the unique equilibrium of the original Email Game—in which HAWK is never played—fails to pass an appropriate evolutionary stability test (Binmore and Samuelson [7]).

6 The Evolution of Conventions

My own view is that to focus on the knowledge requirements for an operational convention is to lose track of what is most important. This isn't to say that what the players in a game may or may not know doesn't matter, but that knowledge issues are secondary. Conventions can sometimes be sustained without anyone knowing anything at all in the formal sense required by current theories of knowledge.

For example, the songs that certain species of birds sing is a cultural phenomenon. Young birds learn to sing complicated arrangements of notes by listening to the songs of experienced birds. It matters a lot to them what song they sing, because the songs are used as a coordinating device in deciding who mates with whom. But the birds don't "know" any of this. Nor I think do humans when they operate most of the conventions woven into our social contracts. As Hume [11] observes, most conventions arise gradually and acquire force by a slow progression. Or, as we would say nowadays, they are the product of a largely unconscious process of cultural evolution.

What is fair? An experiment on the smoothed Nash Demand Game that I ran with some colleagues at the University of Michigan may perhaps serve to illustrate the evolutive attitude to conventions that I favor (Binmore *et al* [8, 5]).

The feasible set in the experiment is shown in Figure 3, with serious money substituting for utility. The exact Nash equilibria correspond to points on the thickened line.¹⁰ The letters *E* and *U* refer to the egalitarian and utilitarian outcomes. The egalitarian outcome is what one gets by applying Rawls' [19]

¹⁰The exact Nash equilibria don't all approximate the Nash bargaining solution *N* because our computer implementation didn't allow the players to vary their demands continuously.

difference principle in this context. The utilitarian outcome is the point in X where the the sum of the players' payoffs is largest (Harsanyi [10]). The letter N corresponds to the Nash bargaining solution (Nash [17]). The letter K refers to an alternative bargaining solution proposed by Kalai and Smorodinsky [12].

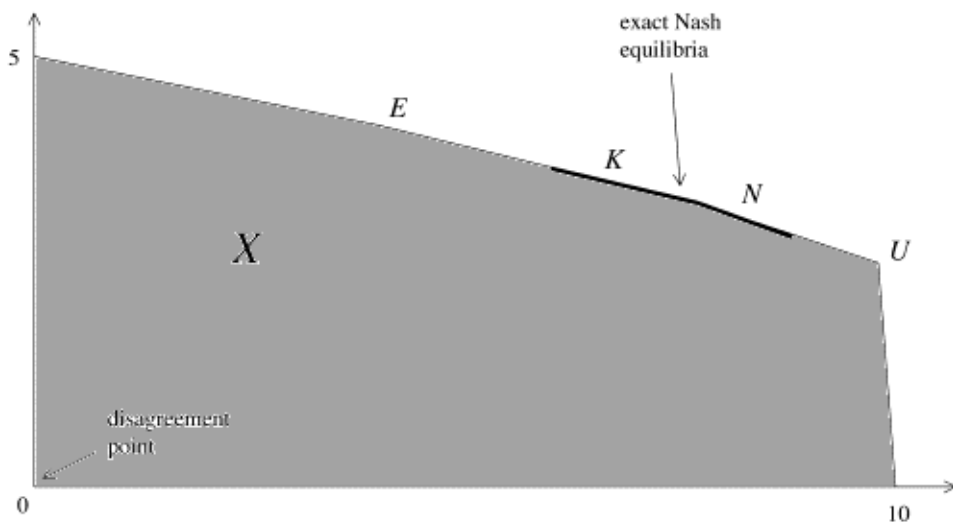


Figure 3: What is fair in the Nash Demand Game?

The experiment began with ten trials in which different groups of subjects knowingly played against robots programmed to converge on one of the possible focal points E , N , K , and U . This conditioning phase proved adequate to coordinate the play of a group on whichever of the four focal points we chose. The conditioning phase was followed by thirty trials in which the subjects played against randomly chosen human opponents from the same group. The results were unambiguous. Subjects started out playing as they had been conditioned, but each group ended up at an exact Nash equilibrium.

It is striking that the different conventions that evolved in the experiment selected only exact Nash equilibria, even though some groups were initially conditioned on the egalitarian and utilitarian solutions, which were both approximate equilibria from which the players would have no incentive to deviate if they neglected amounts of less than one dime.

In the computerized debriefing that followed their session in the laboratory, subjects showed a strong tendency to assert that the convention that evolved in

their own group was the fair outcome of the game. But different groups found their way to different exact equilibria. Indeed, for each exact equilibrium of our smoothed demand game, there was some group willing to say that this was near the fair outcome of the game!

I think the results exemplify David Hume's view of how conventions work. The subjects in each experimental group behaved like the citizens of a minisociety in which a fairness norm evolved over time as an equilibrium selection device.

References

- [1] R. Aumann. Agreeing to disagree. *Annals of Statistics*, 4:1236–1239, 1976.
- [2] K. Binmore. Modeling rational players, I and II. *Economics and Philosophy*, 3 and 4:179–214 and 9–55, 1987.
- [3] K. Binmore. *Playing Fair: Game Theory and the Social Contract I*. MIT Press, Cambridge, MA, 1994.
- [4] K. Binmore. *Natural Justice*. Oxford University Press, New York, 2005.
- [5] K. Binmore. *Does Game Theory Work? The Bargaining Challenge*. MIT Press, Boston, 2006.
- [6] K. Binmore. *Playing for Real*. Oxford University Press, New York, 2006.
- [7] K. Binmore and L. Samuelson. Coordinated action in the electronic mail game. *Games and Economic Behavior*, 35:6–30, 2001.
- [8] K. Binmore, J. Swierzbinski, S. Hsu, and C. Proulx. Focal points and bargaining. *International Journal of Game Theory*, 22:381–409, 1993.
- [9] J. Y. Halpern. Using reasoning about knowledge to analyse distributed systems. *Annual Review of Computer Science*, 2:37–68, 1987.
- [10] J. Harsanyi. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press, Cambridge, 1977.
- [11] 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).

- [12] E. Kalai and M. Smorodinsky. Other solutions to Nash's bargaining problem. *Econometrica*, 45:1623–1630, 1975.
- [13] D. Lewis. *Conventions: A Philosophical Study*. Harvard University Press, Cambridge, MA, 1969.
- [14] J. Maynard Smith. *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, 1982.
- [15] D. Monderer and D. Samet. Approximating common knowledge with common beliefs. *Games and Economic Behavior*, 1, 1989.
- [16] R. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, Cambridge, MA, 1991.
- [17] J. Nash. The bargaining problem. *Econometrica*, 18:155–162, 1950.
- [18] J. Nash. Non-cooperative games. *Annals of Mathematics*, 54:286–295, 1951.
- [19] J. Rawls. *A Theory of Justice*. Oxford University Press, Oxford, 1972.
- [20] J.-J. Rousseau. The inequality of man. In G. Cole, editor, *Rousseau's Social Contract and Discourses*, pages 157–246. J. M. Dent, London, 1913. (First published 1755).
- [21] A. Rubinstein. The electronic mail game: Strategic behavior under “almost common knowledge”. *American Economic Review*, 70:385–391, 1989.
- [22] T. Schelling. *The Strategy of Conflict*. Harvard University Press, Cambridge, MA, 1960.
- [23] B. Skyrms. *Evolution of the Social Contract*. Cambridge University Press, Cambridge, 1996.
- [24] B. Skyrms. *The Stag Hunt and the Evolution of the Social Structure*. Cambridge University Press, Cambridge, 2003.
- [25] C. Turnbull. *The Mountain People*. Touchstone, New York, 1972.
- [26] J. Von Neumann and O. Morgenstern. *The Theory of Games and Economic Behavior*. Princeton University Press, Princeton, 1944.