Is Language a Game?

JOSEPH HEATH
University of Toronto
Toronto, ON
Canada M5S 1A1

Recent developments in game theory have shown that the mathematical models of action so widely admired in the study of economics are in fact only particular instantiations of a more general theoretical framework. In the same way that Aristotelian logic was 'translated' into the more general and expressive language of predicate logic, the basic action-theoretic underpinnings of modern economics have now been articulated within the more comprehensive language of game theory. But precisely because of its greater generality and expressive power, game theory has again revived the temptation to apply formal models of action to every domain of social life. This movement has been fuelled by some notable successes. Game theory has provided useful insights into the logic of collective action in the theory of public goods, and strategic models of voting have illustrated important aspects of institutional decision-making. But this extension of formal models into every area of social interaction has also encountered significant difficulties, despite the fact that contemporary decision theory has weakened its basic assumptions to the point where it teeters constantly on the brink of vacuity.

I believe that these difficulties are not incidental. In this paper, I will attempt to show that an important class of rational social actions cannot be modelled game-theoretically. This is because standard non-cooperative game theory is unable, in principle, to model speech acts, and is therefore unable to specify what is rational about linguistically mediated interactions. At the same time, any attempt to modify the non-cooperative model to include communication as a special type of action generates only paradox and indeterminacy. This suggests that game theory is unsuited, in principle, to serve as a general theory of rational action. If this argument is correct, it would have important methodological implications for every branch of the social sciences.
The term 'general theory of action' refers to the program initiated by Talcott Parsons, aimed at producing an action-theoretic model of dynamic processes that could be applied in every domain of social-scientific inquiry. Such a general theory would serve as a bridge between the disciplines, providing a conceptual framework that would allow results in one field to be translated into those of another. In this paper I am not interested in challenging any aspect of the particular subdiscipline of mathematics known as game theory; I simply want to argue that it cannot be applied in all contexts. In order to emphasize this, I will use the term 'rational choice theory' to refer to the attempt to use game theory as a general theory of action.

But since criticisms of rational choice theory are heard so often, I would like to begin by mentioning two types of objections that I will not be raising. The first, and most common, stems from a general failure to appreciate how weak the basic assumptions of preference-based decision theory are. In particular, many critics confuse the notion of a von Neumann-Morgenstern utility function with the more familiar Benthamite concept of utility, and so understand rational choice theorists to be committed to some dubious assumptions about human psychology and motivation, e.g. that all action is self-interested. Although versions of this basic objection have been advanced without such errors, in this paper I am also not interested in challenging any aspect of the decision-theoretic foundations of game theory.

The second, more important, set of objections involves criticism of the 'unrealistic' aspects of game theory models, mainly the demanding epistemic conditions of Nash equilibria, the sophisticated computational skills attributed to players, and the need for randomized strategies. These concerns are translated into the claim that such a theory of rational action cannot suffice as a general theory of action because people usually do not behave rationally in this precise sense. Thus an adequate general theory would have to accommodate 'psychological' aspects of action, e.g. it is often claimed that 'norm-governed' actions should be classified separately as a nonrational action type. The problem with these sorts of 'empirical' criticisms is that they do not challenge the status of rational choice theory as a normative model of rationality. Thus they make the crucial concession of surrendering rationality to the homo economicus.

My central claim in this paper is somewhat stronger, in that I am not only claiming that game theory cannot serve as a general theory of action, but that it cannot even serve as a general theory of rational action. (If game theory cannot model communicatively mediated interactions, then it would be reasonable to suppose that this sort of action is potentially rational in some other sense.) Game theorists have long been divided over the issue of whether communication should be represented 'inside' or 'outside' the game, i.e., whether speech acts can or cannot be modelled. But the rational choice theorist, because she is committed to the claim that game theory is an adequate general theory of action, must subscribe to the view that all communication can be represented 'inside' the game. In this paper, I will draw on both philosophical and game-theoretic literature to demonstrate the implausibility of this view.

In section I, I outline the basic rational choice agenda and offer a brief characterization of the model. In section II, I present a non-cooperative game-theoretic model of communication. I do so by formalizing David Lewis's model of a signalling system, presenting it as a standard Bayesian sender-receiver game. In section III, I criticize this model, and attempt to show that its shortcomings stem from intrinsic limitations of the game-theoretic approach. In section IV, I present some of the reasons why game theorists have been unwilling to abandon the non-cooperative framework and introduce speech acts as a special class of rational actions. In section V I conclude by pointing out some of the major trouble spots of rational action theory that could be profitably addressed by introducing the assumption that there are not only different types of rational action, but also different types of action rationality.

I Game Theory and Communication

A 'non-cooperative' game is normally defined as one in which 'each participant acts independently, without collaboration or communication with any of the others.' This imposes two restrictions: players are unable to commit themselves to a particular course of action, and players are unable to communicate to each other their intentions or observations. A

---


'cooperative' game is one which includes either or both of these two elements.

Given this distinction, it is perhaps surprising to discover that virtually all of the standard applications of game theory in the social sciences use only non-cooperative models. In fact, when most people refer to 'game theory,' it is actually 'non-cooperative game theory' that they have in mind. But how useful could a model of social interaction be that excludes communication? Certainly if game theory is to provide a general theory of rational action it must be able to supply a game-theoretic account of all 'rational' activities. This will include not only consumption and voting decisions, but also such paradigmatically rational activities as compiling data in a lab, debating economic policy, and writing a journal article on rational choice theory. The exclusion of communication from standard game theory models would therefore appear to limit their applicability rather severely.

In order to maintain the claim to generality, the rational choice theorist must find some way around this limitation. There are two basic strategies for doing this.

The Nash Program: The more ambitious rational choice claim (following the general program staked out by John Nash) is that non-cooperative game theory can serve as a general theory of rational action because non-cooperative foundations can be supplied for all cooperative games. According to this view, there is no difference in principle between games with and without communication. A speech act is simply one more 'move' in a non-cooperative game. Any action can serve to transmit information by virtue of what other players are able to infer about the actor intentions on the basis of the move made. The fact that such an act may involve speech changes nothing. The ban on communication amounts only to the claim that no action taken by any player be able, through an intrinsic property (like an exogenously determined linguistic meaning), to convey information to another player.

Cooperative Game Theory: The fall-back position, should the Nash program fail, is to grant that cooperative games cannot be reduced to non-cooperative ones, but to maintain that cooperative game theory can serve as the general theory of action. Non-cooperative games would then be treated as an interesting class of games within a more general framework. This claim is less ambitious, but technically more problematic. Unlike non-cooperative game theory, which has a single generally accepted interpretation and a very robust solution concept, cooperative game theory is something of a mess. And since there is no obvious reason to believe that any of the results obtained in non-cooperative models will hold true in cooperative ones, the rational choice theorist must show that the adoption of a cooperative framework would not be unduly disruptive.

Most rational choice theorists operate under the assumption that the Nash program can be (or has been) carried out. In the following two sections, I show that it cannot. I then argue in section IV that cooperative game theory does not offer any viable recourse.

The Nash program requires that there be one generic type of action. The 'meaning' of any such action, if it has any, is constituted by the inferences that other players are able to draw from it using standard strategic reasoning. Thus the meaning of any speech act must be determined endogenously in every game, i.e., one cannot assume that meanings are fixed prior to individual acts of communication. (If meanings were determined exogenously, then one would have to model meaningful and non-meaningful actions differently, which by definition would transform the model into a cooperative one.) The result is a somewhat peculiar refraction of the Wittgensteinian doctrine of meaning as use. To know the meaning of an utterance, one must examine its role in the language game. And the 'game' in question is not constituted by any suspicious mental or semantic entities, but rather a set of practical,

---

5 The exceptions are analyses that employ axiomatic bargaining models or assume coalitions. Both of these techniques tend to be avoided because it is generally acknowledged that they lack microfoundations.

6 For example, standard textbooks on game theory like Drew Fudenberg and Jean Tirole, Game Theory (Cambridge, MA: The MIT Press 1991), do not deal with cooperative game theory at all.

7 Nash, 295


9 Jalal Hintikka's 'game-theoretic semantics' are not relevant to this discussion, since he does not try to model communication as a move within a game. Instead, he uses game theory to represent the verification conditions for certain types of sentences. See, for instance, 'Language-Games for Quantifiers' in Logic, Language-Games and Information (Oxford: Clarendon Press 1973).
goal-directed actions (or, more precisely, by a decision tree and a set of payoff vectors).

In evaluating the Nash program, the key question is whether this particular characterization of the social interaction game is rich enough to capture the phenomenon of meaningful speech. Since there is a widespread tendency to use game-theoretic models in very imprecise and metaphorical ways, it is important to begin with a careful review of the modeling tools that are available within the non-cooperative framework. This involves, above all, understanding the relationship between game theory and decision theory.

The basic instrumental model of action, along with the core conception of rationality expressed in the expected utility maximization theorem, is drawn from the realm of decision theory. Decision theory deals with the logic of choice in cases where the actor’s environment does not include any other rational actors (qua rational actors). This allows the agent to calculate the effects of her actions while treating the course of events in the environment as a given, or at least independent of her own deliberations. Because the probability of various states of nature obtaining can be described by a set of statistical assumptions, the external world can be represented as a set of parameters that constrain the agent’s choice.

Decision theory relies upon a very parsimonious ontology. This consists of states, outcomes, actions, and beliefs: a state is a description of the world, that may or may not obtain; an outcome is something that can happen, as it relates to the actor; an action is a function that maps an outcome onto each possible state; and an agent’s beliefs are a set of subjective probability distributions over states. The agent’s preferences can be given as a rank ordering on a set of lotteries that assign different probabilities to various outcomes. With a few additional assumptions a cardinal utility function for the agent can be derived from this preference relation. Rationality consists simply in choosing the action that will give the highest-ranked lottery, given the agent’s beliefs about the prevailing state (when this is represented in a utility function, it shows up as maximizing expected utility).

When the agent interacts with other rational agents, the problem is not so simple, because agents will anticipate each other’s actions, anticipate each other’s anticipations, etc., and develop their plans accordingly. In order to deal with this complication, game theory retains the basic teleological model of action from decision theory, but expands the role of agent beliefs. When dealing with other rational agents, it is impossible to determine the probability of a particular state without taking into consideration how these agents will act. But for an agent to anticipate their actions (in order to form adequate beliefs about the state), she must determine the beliefs of the other players. This in turn will require determining their beliefs about her beliefs, their beliefs about her beliefs about their beliefs, and so on. For a long time, it was thought that this led automatically to an infinite regress. Since the state is no longer given parametrically, the agent must solve for two variables simultaneously. Not only must she decide which action is optimal, she must also determine which state will obtain. The problem is that which state will obtain depends upon which actions are optimal, and which actions are optimal depends upon which state will obtain.12

Game theory was developed when it was discovered that a strategic equilibrium provides a solution to this problem, because it terminates the regress of anticipations and allows a stable set of beliefs to emerge for all players. For each player, strategic reasoning operates by working through the cycles of anticipation in such a way as to turn other players’ actions (which have not yet been planned or performed), into events that will occur with specific probabilities. Once reduced in this way, each player’s problem can then be handled using the standard techniques of decision making under risk. A game-theoretic ‘solution’ is therefore an equilibrium of beliefs, and game theory is a general mechanism for determining beliefs about the relative probabilities of various states.

The view that game theory provides a general theory of rational action involves two separate claims. First, it must be shown that the components that go into building up various types of important rational social interactions can be represented using game-theoretic tools, i.e., that it is possible to characterize any particular problem exhaustively in terms of preferences, actions, outcomes, etc. It is also no less important to then show that these models, once constructed, have solutions.

12 As John von Neumann and Oskar Morgenstern put it: 'Every participant can determine the variables which describe his own actions, but not those of the others. Nevertheless, those “alien” variables cannot, from his point of view, be described by statistical assumptions. This is because the others are guided, just as he himself, by rational principles' (The Theory of Games and Economic Behavior, 2nd ed. [Princeton: Princeton University Press 1947], 11). Thus the traditional “Robinson Crusoe” model of instrumental rationality “is of much more limited value to economic theory than has been assumed even by the most radical critics” (12).

---


11 The clearest presentation of this derivation remains R. Duncan Luce and Howard Raiffa, Games and Decisions (New York: John Wiley & Sons 1957), 12-39.
When it comes to fulfilling the Nash program, this second step is unproblematic, since general existence theorems have been proven for all of the major non-cooperative solution concepts. It is less obvious, however, how various features typical of linguistically mediated interactions should be modelled. One of the usual goals of communication is to affect another person's state of mind, but it is not clear whether this state of mind should be characterized as a belief or an outcome. In order to see how this question is addressed, it is necessary to examine briefly a few relevant developments in the philosophy of language.

The first modern presentation of the instrumental model of rational action is normally attributed to Hobbes. It is less commonly noted that Hobbes was the progenitor of an enormously influential current of opinion in the philosophy of language. This is not coincidental. Hobbes was able to present his basic model of action as a comprehensive one in part because he did not think linguistic interaction presented any exceptions. And it is precisely because of his views on language that he was able to maintain this position.

Hobbes subscribed to a simple causal-nominalist account of meaning, based on what Richard Rorty calls the 'idea idea.' Words are tags that we attach to ideas, and the meaning of the word is the idea that is associated with it in the mind. This means that speech acts are a case of perfectly ordinary teleological interventions in the world. By pricking you with a pin, I cause you to start; by saying 'horse' I cause the idea 'horse' to appear before your mind. This view fits comfortably with the instrumental model of rational action. The Hobbesian account of communication does not even require any game theory, just decision theory. The heater's state of mind is an outcome and the meaning association that establishes the causal connection between word and idea is a state of nature.

This view is now largely discredited, although variations on it persist. One of the more interesting objections that has been raised to this and other causal-associationist accounts involves the failure to recognize the role of what is called the 'Gricean mechanism' in linguistic understanding. In an influential paper, H.P. Grice argued that recognition of an intention to communicate is required in order for the linguistic communication to be successful. Grice distinguishes between what he calls natural and nonnatural meaning. He illustrates the distinction using cases like the following: A person wants to suggest to a friend that her spouse has been unfaithful. He can (A) show her a photograph of the inappropriate conduct, or (B) draw a picture of the same; Herod wants to inform Salome of the death of John the Baptist. He can (A) show her the head on a charger, or (B) say to her 'John the Baptist is dead.' In each of these examples, one individual intends to inform the other of something. But the difference between the A and B cases is that in the A cases recognition of this intention by the hearer is irrelevant to the success of the action, while in the B cases it is not. Thus the connection between the photograph or severed head and their 'meanings' is in a certain sense natural. But in the B cases, communication would fail if the hearer interpreted the drawing as simply doodling, or the spoken words as just noises.

Grice went on to make the more dubious claim that the specific meaning of the utterance was also determined by the intention. But we do not have to accept this to see that the recognition of an intention to communicate is necessary in order for an action to be construed as meaningful. The hearer is not simply acted upon by the speaker, but participates actively in the exchange. Thus the mental state produced by communication cannot be an outcome, i.e., the action cannot be modelled decision-theoretically. The nonnatural quality of linguistic meaning gives it no place in the state-outcome domain of the game-theoretic ontology.

For these reasons, game theorists have followed David Lewis in attempting to represent communication as a process that produces beliefs rather than outcomes. Lewis's model of meaning conventions basically modifies associationism in order to build in a Gricean mechanism. He argues that communicative understanding is an equilibrium in a multiple-equilibrium coordination game. The recognition of an intention to

---

13 See Fudenberg and Tirole, *Game Theory*, 29-36, or Roger Myerson, *Game Theory: Analysis of Conflict* (Cambridge, MA: Harvard University Press 1991), 136-40. Note that the only general solution concepts are backward induction based, i.e., they start with an outcome and reason back to the required actions. This can be very counter-intuitive, and should therefore always be kept in mind.


16 I think this should be non-controversial. Nevertheless, it should be noted that a prima facie consequence of this claim is that Donald Davidson's program for a unified theory of meaning and action cannot be carried through as he intends, since he (somewhat inexplicably) limits the action-theoretic component to simple decision theory. See 'The Structure and Content of Truth,' *The Journal of Philosophy* 87 (1990) 279-328.

communicate is modelled as part of the normal cycle of anticipations that hold together these equilibria. In this way, the Gricean mechanism is shown to be nothing other than a special case of the standard strategic reasoning used in all contexts of social interaction.

In this model, the hearer's state of mind is not treated as an outcome that can be achieved through direct teleological intervention. Instead, information transmission is represented as the development of rational beliefs by the hearer in the course of a game in which all players are 'ultimately' pursuing non-linguistic goals. Since the set of beliefs that serve as the strategic equilibrium of the game also reflect the true state of the world, the hearer is able, through strategic reasoning and observation of the speaker's action, to determine the effective equilibrium, and thereby determine the state of the world. So, unlike the Hobbesian model, communication here is intrinsically game-theoretic. On the other hand, associationism is preserved as a means of solving the problem of equilibrium selection. Communication does not determine the equilibrium reached, because the meaning of all statements is determined endogenously by the equilibrium itself. Instead, an equilibrium is selected on the basis of its salience, and this salience is generated by the associations formed by habit. Associationism thereby provides the determinacy of outcome that standard game theory lacks.

Lewis did not intend his model to be a comprehensive account of language use. But I consider it a reasonable supposition that any rational choice approach to the study of communication would have to be developed along similar lines. I will therefore present a Lewis-style model of communication, using a contemporary game-theoretic formalization.

II Sender-Receiver Games

There are two general methods for representing games: the extensive and the normal (or strategic) form. Although the extensive form is less familiar in the philosophical literature, I will be using it here because it represents explicitly certain features that are either lost or obscured in the normal form. The extensive form of the familiar prisoners' dilemma game is shown in Figure 1.

The extensive form of a game of complete information can be described by a sextuple: \( \Gamma = (K,P,H,C,p,u) \).

- \( K \) represents the game tree, as a finite set of nodes ordered by a binary precedence relation. The origin is represented graphically by an empty circle, decision nodes are represented by dark circles, and terminal nodes by the associated payoff vector.
- \( P \) partitions the set of non-terminal nodes into player sets, determining which decision nodes 'belong' to which player, i.e., whose move it is at each point. \( P = \{P_0, ..., P_n\} \), where player 0 represents 'nature,' who is responsible for random decisions in the game. In Figure 1, the origin belongs to player 1, while the other two decision nodes belong to player 2.
- \( H \) is the information partition, which divides each player's decision nodes into information sets. In the course of the game, players do not know the node they have reached, only the information set. This is represented by a dashed line connecting all the nodes contained in the information set (some may be singleton sets). In Figure 1, the dashed line indicates that when player 2 is called upon to move, she does not know whether player 1 has chosen C or D.
- \( C \) is the choice partition, which specifies the options available at each information set. Each different choice is given a unique move label. In Figure 1, these are \( \{C,D,c,d\} \).
- \( p \) specifies the probability of each random event in the game, i.e., of each choice for each node belonging to player 0, nature.

Finally, the payoff function \( u \) assigns a von Neumann-Morgenstern utility to each terminal node. In the diagram, this is shown as a vector of payoffs (player 1 first, player 2 second, etc.).

All of this information is taken to be something that everyone knows, and that everyone knows that everyone knows, etc. The only form of 'ignorance' present involves players not necessarily knowing what other

---

18 This is the 'psychological' solution developed by Thomas Schelling in *The Strategy of Conflict* (Cambridge, MA: Harvard University Press 1960).

19 Most obviously, the game-theoretic segment does not handle compositionality.
Players have done in the course of play, which is represented through information sets (analysis is normally restricted to games of ‘perfect recall’ where players always know their own past moves). In the standard prisoners’ dilemma shown in Figure 1, the fact that player 2 cannot distinguish her two decision nodes is used to represent both players moving simultaneously. 20

The most familiar solution concept is that of Nash equilibrium, which is a strategy profile (a set of strategies, one for each player), in which no player can gain by unilaterally changing strategies. Figure 1 has only one Nash equilibrium, \([D],[d]\). However, this example obscures a very important detail. \([d]\) is player 2’s best response, no matter what player 1 does, and so she is actually not concerned which node in her information set has been reached. But there are many instances in which the player’s beliefs about where she is in an information set will determine the strategy she pursues. Consider Figure 2:

![Figure 2](image)

This game has two pure strategy Nash equilibria, \([C],[c]\) and \([D],[d]\). Notice here that Player 2’s willingness to do her part of the first equilibrium depends upon her believing that she is at the top node of her information set. Because of this, David Kreps and Robert Wilson have argued that any equilibrium should consist not of a set of strategies, but rather of a set of assessments, defined as ‘a pair \((\mu,\pi)\) consisting of a system of beliefs and a strategy \(\pi\).’ 21 Belief probabilities for each node are indicated in the game diagram between angle brackets \((<\cdot\cdot\cdot>)\). In Figure 2, the \([C],[c]\) equilibrium requires that \(\alpha > \frac{1}{2}\). This representation renders explicit a dimension of the equilibrium concept that is present yet unstated in the standard formulation of Nash equilibrium.

In order to be in equilibrium, the assessments must contain a belief system that is consistent with the strategies being played. This is a somewhat murky doctrine, but the most important idea is that beliefs must be formed using Bayes’s rule, i.e., the belief that I am at node \(x\) of information set \(h\) should be determined by

\[
\mu(x) = \frac{P^\pi(x)}{\sum_{y \in h} P^\pi(y)}
\]

where \(P^\pi(x)\) represents the probability of being at \(x\) given strategy \(\pi\), and the denominator gives the probability of being anywhere in the information set \(h\), given strategy \(\pi\). Bayes’s rule is important, because it is the only mechanism through which players are able to update their beliefs in the course of the game. Unlike as it might seem, this means that the meaning of sentences in rational choice models of communication will have to be inferable using Bayes’s rule.

However, the model elaborated so far is not yet very useful, because everything in the game is common knowledge. In order to have interesting communication, it must be the case that at least one player knows something that at least one other does not. This requires a somewhat special representation, since in a standard game it is assumed that all of the information available to the theorist about the game is both available and common knowledge among the players. To represent games where some players have private information, known as games of incomplete information, John Harsanyi introduced the ‘Bayesian’ form. In such games, either payoff functions, available choices, or information about the game is not shared. His formulation permits the translation of games of incomplete information into ‘Bayes-equivalent’ games of complete

---

20 In this paper, I follow the convention of using male pronouns to refer to odd players and female pronouns for even.

information by representing players with different amounts of information as different 'types'.

A type is a summary of everything the player knows about the game that is not common knowledge. In order to make this situation mathematically tractable, the type of each player is taken to be determined by a historical chance node, and the complete set of types, along with the prior probability distribution over this set, is taken to be common knowledge. (Note the implication: every player must know, in advance, the probability of every possible state of the world that could in any way affect the game. This is an heroic assumption, but there seems to be no way around it, since without it every Bayesian game gets into an infinite regress of anticipation.)

An extensive form Bayesian game is represented by the septuple: \( \Gamma = (K,P,T,H,C,p,u) \). This is the same as above, except \( T \) has been added as the set of types, and the utility function is modified to reward players according to type. The probability function \( p \), which specifies the random events in the game, determines the type of each player through a 'nature' move at the beginning of the game.

This sort of game can now be used to model communication. The most general type of language game is called a sender-receiver or signalling game. Player 1, the sender (S), has private information, i.e., a type; and player 2, the receiver (R), has none. S selects an action, called the 'message,' after which R chooses an action that determines the payoff for both players.

Consider the following example. Two builders are busy at work. S is building a wall, while R is passing him the materials he needs. R cannot see the wall, and so only S knows what materials are needed at what time. Suppose there are two types of materials being used: blocks and slabs. This gives us two types of S: the need-a-block type and the need-a-slab type. S sends a message to R. After receiving the message, R passes S either a block or a slab. If R passes a block to the need-a-block type S, or a slab to the need-a-slab type S, then they have successfully coordinated and are both happy. This game is illustrated in Figure 3.

24 Lewis uses an example that involves Paul Revere hanging lanterns in the belfry (Conventional, 122-3). The structure of the game shown in Figure 3 is the same, except that it has only two messages instead of three. We could easily add more types to the above game, e.g., pillars and beams, but it would make Figure 3 even less intelligible.

25 This type of game tree diagram can be hard to read at first. The origin is in the center, and nature starts the game by moving up or down, thereby determining player 1's type. From either 1.a or 1.b, player 1 then moves left by shouting 'block' or moves right by shouting 'slab.' If Player 2 hears 'block,' then she is at information set 2.a, and is at 2.b if she hears 'slab.' She then passes either a block or a slab, ending the game.

---

22 John Harsanyi, 'Games with Incomplete Information Played by "Bayesian" Players', 3 parts, Management Science 14 (1968) 159-82, 220-34, 486-502. The idea, roughly, is to take the single game of incomplete information and represent it as a set of different games of complete information, where not all players know which game is being played. In a similar way, one might represent the vagueness of a predicate as a probability distribution over the set of acceptable sharpenings.

a slab. If a block is handed over to the need-a-block type, or a slab to the need-a-slab type, both players receive a payoff of 1, otherwise they each get 0.

In this game, there are two pure strategy equilibria (informally): \([S \text{ says 'block' at } 1.a, S \text{ says 'slab' at } 1.b, R \text{ passes block at } 2.a, R \text{ passes slab at } 2.b] \), and \([S \text{ says 'block' at } 1.b, S \text{ says 'slab' at } 1.a, R \text{ passes block at } 2.b, R \text{ passes slab at } 2.a] \). Both of these equilibria yield a payoff of 1 to each player. There is also a plethora of mixed-strategy equilibria, where both players randomize in various ways over their pure strategies. These all yield a payoff of \(1/2\) to each.

Consider the first equilibrium. Given what the other player is doing, neither has an incentive to deviate. If \(R\) is passing a block whenever she receives the message 'block', and a slab whenever she hears 'slab,' then obviously \(S\) has no incentive to shout 'slab' when he in fact needs a block. And if \(S\) is shouting 'block' only when he needs a block, etc., then \(R\) also has no incentive to deviate. Knowing this, and using Bayes's rule, it is rational for \(R\) to assign \(\alpha = 1\) when she hears the message 'block,' and \(\beta = 0\) when she hears 'slab,' because these are the only beliefs consistent with actions \(S\) would perform. Thus the Nash equilibrium being played requires that \(R\) form correct beliefs about which node in her information set she has reached, and by implication, \(S\)'s type. It is therefore no stretch of the imagination to say that the meaning of the messages 'block' and 'slab' are the types that they reveal in the equilibrium. By sending a message that allows the hearer to infer the true state of the world, the speaker has effectively communicated information to the hearer. The meaning of the expression is simply this information content.²⁶

We now have a simple way of characterizing the second equilibrium. It is the same as the first, except that the meaning of the messages 'block' and 'slab' have been reversed. This is nothing other than the conventionality of linguistic meaning. Both players are indifferent between the two equilibria, since their payoffs remain the same. They do not care which sounds have which meanings, as long as they are able to communicate effectively. Since there are no rational grounds for preferring one of these equilibria over the other, Lewis argues that the particular 'meanings' messages have will be determined merely by force of habit. Once players have observed a certain number of plays of the game where the same equilibrium is selected, they will tend to stick to that equilibrium.²⁷ In the same way, once we have met in a particular café a couple of times, we will tend to return to the same café if we want to meet again. Therefore, the 'psychological solution' to equilibrium selection, in this case associationism, retains its central role in communicative interaction.

III Problems

There can be no doubt that this model of communication is extremely clever. Whenever the sender is divided into types, the receiver will have information sets in which different nodes correspond to different types. By determining which node she has reached, using Bayes’s rule, \(R\) is able to infer \(S\)'s type, and thereby acquire its information content. But the problems with it are somewhat obvious, and have been noted by several authors. Here I would like to bring some of this analysis together, in order to show that these problems stem from intrinsic limitations of the non-cooperative model.

The first problem relates to equilibrium selection. In the building game, there is a plethora of mixed strategy equilibria in which both players randomize over their pure strategies (e.g., when \(R\) assigns a probability of \(1/2\) to each action, \(S\) can select any randomization over messages, as long as it is the same for both types). These equilibria are counterintuitive, partly because they are suboptimal. Nevertheless, if \(R\) is flipping a coin to decide what materials to hand over, then \(S\) does not care what message he sends. And if \(S\) is flipping a coin in order to determine whether to say 'block' or 'slab,' \(R\) is completely indifferent between her two actions. Thus we have the famous 'babbling equilibria,' in which each player randomizes over his possible statements, and all other players ignore him. In this situation, all messages are meaningless.

These equilibria are extremely implausible, but there appears to be no way of eliminating them through refinements.²⁸ Lewis avoids the problem by legislating mixed strategies out of the model, an option that is not

---

²⁶ Lewis adds a clever mechanism for determining whether the utterance is an assertion or an imperative, which need not concern us here (Convention, 144).

²⁷ Russell Hardin in his Collective Action (Baltimore: Johns Hopkins University Press 1980), 163, misunderstands Lewis to be suggesting that linguistic conventions arise in a repeated game (in which case the folk theorem would apply and each game would have an infinite number of equilibria). Lewis's claim is only that past plays of the game determine the salient outcome, they are not part of the equilibrium. This is what allows us to learn the meaning of a message by watching others play the language game.

²⁸ Myerson, Game Theory, 373-4
available to the rational choice theorist, since it undermines the generality of the Nash solution concept. In any case, it is an artificial solution, one which ignores the deeper problem. A babbling equilibrium seems to correspond to the case where there is no effective 'Gricean mechanism.' Words are treated like noises. And since the messages do not directly affect payoffs, i.e., they do not accomplish anything in the world, they are ignored. But the babbling equilibrium is a Nash equilibrium like any other, with exactly the same type of belief supports as the more 'informative' ones. This suggests that game-theoretic reasoning alone does not capture the relevant dimensions of intentionality operative in communicative interaction, and that Lewis's 'higher-order' belief diagrams may well overestimate the determinacy of the Nash solution concept.

A more significant difficulty rests with the informative equilibria. The game-theoretic model makes the meaningfulness of the speech act dependent upon the payoff structure of the game. The unfortunate consequence is that 'the informativeness of the most-informative equilibrium is limited by the degree to which the players' interests coincide.' The only way that R can get any insight into S's intentions is by considering the outcome S is pursuing, along with S's anticipation of the outcome R is pursuing, etc. However, it is not always the case that a player will benefit from having another player anticipate his actions or know his type. In the case of a conflict of interest, S does not have any incentive to inform R of his type, and R, knowing this, has no reason to take any message as indicative of S's type. All messages in the model will therefore be meaningless. Note that this is not a peculiarity of the game that Lewis presents. Vincent P. Crawford and Joel Sobel have established as a completely general limitation result that in the entire class of cheap-talk signalling games 'perfect communication is not to be expected in general unless agents' interests completely coincide, and that once interests diverge by a given, "finite" amount, only no communication is consistent with rational behaviour.'

Suppose that S and R have been working on their wall for a number of days, and are comfortably playing one of the two informative equilibria. Suddenly S discovers that he is to be paid by the hour, not by the job. He no longer has any incentive to build the wall efficiently; in fact, he would like it to take as long as possible. Instead of the game having the payoff structure of a coordination problem, (0,0) and (1,1) for R's two actions, it would now have (0,1) and (1,0), a conflict of interest. With these payoffs, S would benefit from a 'tower of Babel' situation, in which he was no longer able to coordinate his actions with R through a common language. Miraculously, his wishes are self-fulfilling. By simply acquiring this new incentive, all of his speech suddenly becomes meaningless. It is now impossible for R to understand anything S says, because the only equilibrium of this new game is a babbling one. Since S could never benefit from revealing his type, and since R knows this, R will ignore all of S's messages. Now when S calls out 'block' it is just a sound, with no meaning whatsoever.

There is obviously something wrong here. But notice that associationism cannot save the day. One might criticize the example on the grounds that players, having responded to the message 'block' with a certain action for so many days, would retain a residual association even after the game changed. But this is not the role that associationism played in the model. Past instances of communication were used to give certain equilibria salience, not certain sounds meanings. The meaning of the sounds was always inferred from within the equilibrium. In this new game, the two informative equilibria have vanished entirely, and thus no game-theoretic account of language could possibly retrieve them. Linguistic conventions can exist only where there is a coordination problem, and there is no longer a coordination problem here. The new set of payoffs does not permit R to determine anything about S's intentions, and therefore no communication is possible.

This problem usually gets missed because of the tendency to use game-theoretic language metaphorically, in ways that clearly violate strictures of methodological individualism. Lewis says language is a convention, and conventions exist where there is a common interest in achieving coordination. Since it is obviously in our common interest to be able to communicate, meaning conventions must be the solution to a general social coordination problem. The problem is that it may be in our interest, in general and as a group, to have a common language, but game theory deals only with particular situations and individual players. And there are obviously many cases in which it is not in my interest to be able to understand you, e.g., when you are telling me to do something that I do not want to do. It is therefore not adequate to show that social life, in general, has the structure of a coordination problem. It must be shown that on each and every occasion in which any meaningful communication takes place, the interaction has the structure of a coordination problem. This is not just a tall order, it is patently false.

Our normal response to the new building game would be to say that R still understands what S is saying, she is just no longer sure whether or not to believe him. What has clearly gone wrong is that meaning has


30 Crawford and Sobel, 1450
become so closely tied to credibility that it has becomes impossible for anyone to tell an intelligible lie. This is because the non-cooperative model determines a player’s communicative intentions from what we would normally call his ‘ulterior’ motives. This is not incidental. Recall that in the game-theoretic ontology, only outcomes can be the object of an action. Beliefs are developed indirectly. This means that communication can only be achieved indirectly. If your beliefs were the objective of my action, then you would have no way of developing beliefs, because you would have no way of grasping my intentions. You are able to infer my type only if I am pursuing certain objectives in which your state of mind plays at most an instrumental role. My primary objective in speaking to you can never be to have you develop certain beliefs, because this is not an admissible outcome.31 Thus rational agents cannot just talk to each other, or, as Lewis puts it, there can be no idle conversation.32

It would be easy to underestimate the seriousness of these difficulties. According to the rational choice model, there can be no meaningful communication without a significant convergence of interest among the parties, and players cannot pursue strictly communicative objectives. Both of these implications are obviously false, and point to much deeper conceptual confusions. And it is fairly easy to see that neither of these difficulties can be overcome within the framework of non-cooperative game theory without undermining its decision-theoretic foundations. It is therefore reasonable to conclude that important classes of rational social interactions cannot be modelled as non-cooperative games.

IV Cooperative Alternatives

In non-cooperative games, players are able to form beliefs by drawing upon two types of knowledge: knowledge of nature obtained through observation, which allows them to determine the relative likelihood of various states obtaining, and knowledge of rationality obtained through introspection, which allows them to anticipate the actions of others. Suppose we add a new category, knowledge of language (e.g., English), which allows players to assign literal meanings to statements. Now when a player hears the word ‘block’ she knows that it means (in some unanalyzed sense) block. Players can therefore make assertions about their type with the anticipation that they will be understood, even though they may not be believed.

Consider for the moment how this affects our building game. Clearly, a distinction must now be drawn in the model between messages and actions. For simplicity, a cooperative game of this type can be represented with an octuple: \( \Gamma = (K,P,T,H,C,M,p,u) \), which is as above except that M has been added to specify the nodes at which players can send messages (with no restrictions placed on the content of these messages). Instead of having a branch of the tree for each possible message, we can redefine a strategy for this game as a set specifying an action for each node and a message for each message-node, for each possible history of messages received.

This reformulation looks as if it will clean up the building game quite nicely. In the coordination version, it eliminates the ‘babbling equilibria’ as well as the informative one in which the ‘meaning’ of the messages is reversed, leaving only the informative equilibrium in which ‘block’ reveals ‘need-a-block’ and ‘slab’ reveals ‘need-a-slab.’ For the conflict version, the ‘babbling’ mixed strategy equilibrium remains the only solution, and so no effective communication can take place. But it is no longer the case that the sender’s messages are meaningless, they are just not credible. There is nothing peculiar about this, since it appears to be an adequate characterization of the situation in which one party has a constant incentive to lie.

But what sort of solution concept will do this work? In keeping with the instrumental conception of rationality, players presumably send messages for the same reason that they perform actions — in order to maximize their payoffs. Nash equilibrium in non-cooperative games specifies that no agent have any incentive to deviate unilaterally from the equilibrium strategy profile. For the cooperative game, a strategy profile would include both actions and messages, thus an appropriate solution for such a game would have to be one in which no player had an incentive to perform any action or send any message that was not part of the equilibrium strategy profile. A message that is not a part of this

31 Jonathan Bennett has suggested a modification that would allow beliefs to be treated as outcomes in non-cooperative games, ‘The Meaning-Nominalist Approach’, Foundations of Language 10 (1973) 141-68. He expands the category of actions to what he calls doing. These include the ‘act’ of belief-acquisition, and so it becomes possible to treat beliefs as outcomes. But this modification creates enormous tension with the utility-maximizing conception of practical reason. If beliefs were outcomes, they would have to be associated with payoffs for both players. If the agent faced with doing a belief-acquisition did so on the basis of these payoffs it would effectively condone ‘wishful thinking’ and thus spell the end of belief rationality. But if the belief-outcome is still determined belief-rationally, and the payoff is relevant only to the sender, then no longer has any strategy. Without a strategy, there is nothing for R’s beliefs to be consistent with, nothing for S’s strategy to be in equilibrium with, and thus no end to the regress of anticipations.

32 Lewis, 160
equilibrium is standardly referred to as a *neologism*. Thus the solution concept, due to Joseph Farrell, is called *neologism-proof equilibrium*.35

To see how this solution works, consider a version of the building game in which S had some other messages at his disposal. Suppose that he could also call out ‘pillar,’ even though the wall did not require any pillars and the type need-a-pillar occurred with probability zero. In the informative equilibria, the word ‘pillar’ would be a neologism. In order for these equilibria to be neologism-proof, it would have to be the case that S never had any incentive to deviate and send this message.

In the non-cooperative game the meaning of all statements is inferred using Bayes’s rule, and since Bayes’s rule does not place any restrictions on the formation of beliefs in response to probability-zero events, if R were to unexpectedly receive the message ‘pillar’ she could take it to mean anything she wanted. This means that there are always equilibria in which all neologisms are interpreted as meaningless, in which case S does not have any incentive to deviate. Thus the informative equilibria in non-cooperative sender-receiver games are all neologism-proof.34

This does not hold, however, when statements are taken to have literal meanings. Since neologisms will always be understood, only statements that are not credible can be ignored by other players. Thus situations may arise in which players can benefit by deviating from equilibrium play and sending an unexpected message. So in order to develop a strategy, S must determine how R would respond to all of the various messages that he could send. Farrell proposes the following analysis: Take \( m(X) \) to represent a neologism asserting that player S is of type \( t \in X \), where \( X \subset T \). It would be overly credulous of R, upon receiving this message, to infer that \( t \in X \). Instead, she might infer that S is a member of some type that would *prefer* that R believe that \( t \in X \). But then again, she might infer that S is of some type that would prefer that R believe that S is of some type that would prefer that R believe that \( t \in X \), and so on. However, there is a special class of cases in which the set of types who would prefer that R believe \( t \in X \) is just \( X \), i.e. those where S would like R to believe \( m(X) \) just in case it is true. Farrell refers to these as self-signalling neologisms. They have the distinctive property of being *intrinsically credible*.

In the modified building game, ‘pillar’ would not be a credible neologism. But there may be games in which S’s utterance of a neologism would be credible, and so might cause R to change her beliefs, and thereby cause her to change strategies in a way that benefited S. It will therefore be a minimum adequacy criterion of any solution concept for games with exogenously determined meanings that under any equilibrium strategy profile no player have both an available self-signalling neologism and an incentive to use it. Unfortunately for cooperative game theory, it turns out that by this criterion, a wide range of games turn out to have no solution. Consider Figure 4:

![Figure 4](image)

In this game, we can imagine player 1 (S) and player 2 (R) still working on their wall. Nature determines player 1’s type as before, then player 1 sends a message to player 2 (message nodes are indicated by gray circles). Since the messages are in English, they are not represented in the game tree diagram; instead, the turn passes directly to player 2. Upon hearing the message, player 2 has three actions, \( \{b,c,s\} \) which we will interpret as ‘pass a block,’ ‘take a coffee break,’ and ‘pass a slab,’ respectively. For the sake of discussion we can represent player 1’s messages in the following way: anything asserting that he is at node 1.a will be referred to as \( m(a) \), anything asserting 1.b as \( m(b) \), and anything irrelevant, obscure, or equivalent to ‘I’m not telling’ as \( m(t) \).

Player 2’s payoffs are fairly simple. She prefers to hand player 1 the correct materials, but when she is somewhat uncertain about what is needed \( \alpha > \frac{1}{3}, \frac{2}{3} > \alpha > \frac{1}{3} \), i.e. her belief about what materials are needed is

---

33 Farrell, 'Meaning and Credibility in Cheap-Talk Games'

34 The only problem with neologisms is that they multiply the set of equilibria enormously, e.g., R could take ‘pillar’ to reveal type need-a-block, and then there would be all sorts of new equilibria in which S randomized in various ways between ‘block’ and ‘pillar’ every time he was of type need-a-block. But other than creating this annoying proliferation of equilibria, neologisms do not create fundamental problems for any of the traditional non-cooperative solution concepts.
close enough to the prior probability, she would rather call a coffee break. Player 1 wants to get the wall finished, but hates carrying slabs. When a block is needed he wants to receive the block, but failing that he would rather player 2 call a coffee break than hand him a slab. When a slab is required, he is happiest when player 2 calls a coffee break, failing that he would still rather be handed a block than a slab.

This game has no neologism-proof equilibria. Consider first the situation in which R takes all messages to be uninformative and calls a coffee break. Here, m(a) is a self-signalling neologism. However, if R passed a block in response to m(a), she would also pass slabs in response to anything else. Then S would also want to send m(a) when his type was 1 b., in which case it would be better for R just to call a coffee break. Thus for both players, reasoning about this game results in an infinite regress.35

The fact that very simple games like this have no solution seriously impairs the ability of rational choice theorists to claim that cooperative game theory can serve as a general theory of action. The standard response among game theorists has been to add a mediator to the game, who controls the transmission of messages between players.36 Although this move solves the technical problem, it is not available to the rational choice theorist because it makes the model too specific to serve as a general theory of action.

The problem of neologisms, however, is somewhat minor compared to some of the larger unanswered questions that arise within the cooperative framework. In the game considered above, all of the communication that takes place is completely public, i.e., the content of the messages that are sent is common knowledge among players. But if private communication channels are available through which players can reveal their types to each other (without a mediator), then players in games of incomplete information will no longer be able to assign probabilities to each other’s beliefs. This creates a situation exactly like one in which there is no common prior, so the combination of Bayesian and strategic reasoning generates an infinite regress.37 Introducing knowledge of language as a new source of beliefs therefore undermines the standard solution concept for games of incomplete information, suggesting that cooperative game theory will require some non-Bayesian account of belief-updating.

Furthermore, our attention so far has been confined to situations in which players use their available messages only to reveal their type. But as Roger Myerson has observed, the problem becomes more complex when we grant that players can also use messages to announce their intentions, or make requests.38 In a multiple-equilibrium game, a player could use a message node as an opportunity to announce how he intends to play at a later point. This introduces extremely undesirable ‘forward induction’ elements into the game. Since there is no game-theoretic account of equilibrium-selection (only the psychological ‘salience’ theory), it is completely unclear what effect this should have on the other players, and by implication, how the strategy underlying such an announcement should be characterized.

All of this makes cooperative game theory very unattractive as a fall-back position for the rational choice theorist. Earlier I suggested that in order to establish the claim that game theory provides an adequate general theory of action, the rational choice theorist must show that social action can be modelled game-theoretically, and that these models, once constructed, have solutions. When it comes to providing a game-theoretic account of communicative action, non-cooperative models fail on the former count, while cooperative ones fail on the latter.

V Conclusion

Although my argument has so far been entirely negative, my intentions are not limited. The failure to model communication game-theoretically, rather than closing off an avenue of inquiry, opens the door for alternative conceptions of action rationality. Because of the hegemony of the instrumental view, the failures of the rational choice program have often been translated into the claim that ‘rationality’ is not a useful social-scientific concept. In the face of growing skepticism among game theorists about the viability of rationality-based action theory, a broader typology of rational action offers the possibility of an attractive solution to some important outstanding issues. The basic approach, as I see it, would involve developing a concept of communicative action that articulates the underlying rationality of linguistically mediated interac-


37 Myerson, 'Credible Negotiation Statements and Coherent Plans,' 266
I would like to suggest a way of carrying out this project, which naturally I cannot argue for at any length. Given the troubles with cooperative game theory, the non-cooperative model should be accepted as an accurate characterization of the nature, scope and limits of instrumental rationality. The instrumental model draws a distinction between beliefs and outcomes, where the goal of any agent reasoning instrumentally is by definition an outcome. Let us suppose that speech acts have as their goal the production of beliefs. Thus when an agent performs such an act, she cannot be reasoning instrumentally. This appears to leave the door open for alternative specifications of the rationality of the act. There are a number of possibilities here, but one in particular that I consider worthy of serious consideration. Jürgen Habermas has argued that the rationality of a speech act should be characterized solely as a function of the justifiability of its content, e.g., an assertoric speech act is rational if the belief that it asserts is rational, i.e., if the agent has good reasons for believing it. The interesting feature of this proposal is that it makes the rationality of the action a purely cognitive affair.

In this type of model, communicative and instrumental rationality would be conceived of as two pure types, neither of which would occur very often on its own. Strategic action usually occurs in linguistically structured contexts, and communicative actions generally take place in situations where agents are also pursuing other objectives in addition to their communicative goals. Thus everyday interactions would be more than one-dimensional in that both types of rationality-considerations would always be at play. The analysis would be complex, but this merely reflects what microsociological studies have long suggested — the structure of social action is very complex.

The first issue that could be addressed through such an approach would be the problem of social norms and rule-governed behaviour. Without getting into technicalities, the traditional quandary of sociological theorists is something like the following: social action is largely norm-governed, and hence not rational in any obvious sense; at the same time it is highly organized and reflexive. But if it is not rational, then the coherence of norm systems is extremely difficult to explain, since it cannot be achieved through any specifically intentional process. This makes it very tempting to adopt fairly large-scale functionalist commitments, something that those attracted to the action-theoretic framework quite legitimately shy away from. The result is often a sort of theoretical limbo, in which theorists grant an important role to social norms without having any explanation or use for them.

The idea of a non-instrumental form of rational action opens up the possibility that norm-governed actions could be explained as a species of rational action. Again, Habermas has suggested that norm-governed action should be analyzed as a generalization of communicative, rather than instrumental, rationality. Although the connection between linguistic understanding and social norms is not entirely obvious, a successful marriage of the two could help resolve some of the most pressing issues in both sociological theory and moral philosophy.

The second major set of problems that could be fruitfully approached using a broader account of action rationality concerns the indeterminacy of traditional game-theoretic solutions. Since Schelling's seminal work, it has been traditional for rational choice theorists to rely on extra-rational mechanisms to determine equilibrium selection. A theory of communicative rationality, coupled with a rationality-based account of social norms, would provide a clearly superior set of resources for explaining the stability of social orders.

Finally, the attractions of evolutionary game theory, which has been steadily drawing game theorists away from a voluntaristic theory of action, would be considerably diminished. In particular, there is the growing sense that the perfect knowledge attributed to rational players in traditional game theory is not a useful idealization. Not only does it seem reasonable to suppose that players could be uninformed without being irrational, the perfect knowledge assumption eliminates any possibility of understanding the impact of learning processes on social interaction. A theory of communicative action could not only drop this unwanted idealization, but could also facilitate the introduction of a non-Bayesian account of belief formation.


42 This is, as far as I can tell, Jon Elster's current position. See *The Cement of Society* (Cambridge: Cambridge University Press 1989).