[7] In "Anti-Anti-Relativism," in *Relativism: Interpretation and Confrontation*, ed. Michael Krausz, Notre Dame, 1989.
[8] These considerations are elaborations of an argument in Plato's *Theaetetus*, 161 d.e.
[9] See Wittgenstein's *On Certainty, passim.*
[10] The work of J. L. Austin is perhaps the most engaging treatment of this topic.

REFERENCES

Aristotle. *Nicomachean Ethics*. Indianapolis: Bobbs-Merrill, 1962.
Carroll, Lewis. "The Hunting of the Snark," in *The Lewis Carroll Book*, New York: Tudor Publishing, 1944.
Chapman, Graham, *et al. The Complete Monty Python's Flying Circus: All the Words*, Vol. II, New York: Pantheon Books, 1989.
Conway, Gertrude D. *Wittgenstein on Foundations*. Atlantic Highlands: Humanities Press, 1989.
Davidson, Donald. "On the Very Idea of a Conceptual Scheme," in *Language and Reality*.
Fogelin, Robert J. "Wittgenstein and Classical Scepticism," in *The International Philosophical Quarterly*, March, 1981.
_____, *Evidence and Meaning*. New York, Humanities Press, 1967.
Geertz, Clifford. *Relativism: Interpretation and Confrontation,* ed. Michael Krausz, Notre Dame: Notre Dame University Press, 1989.
Kuhn, Thomas. *The Structure of Scientific Revolutions*. Chicago: The University of Chicago Press, 1967.
MacIntyre, Alasdair. *After Virtue*. Notre Dame: Notre Dame University Press, 1981.
_____. "Relativism, Power, and Philosophy," in *Relativism: Interpretation and Confrontation*, ed. Michael Krausz. Notre Dame: Notre Dame University Press, 1989.
_____ *Whose Justice? Which Rationality?* Notre Dame: Notre Dame University Press, 1987.
Nozick, Robert. *Philosophical Explanations*. Cambridge, Massachusetts: Harvard University Press, 1981.
Trigg, Roger. *Reason and Commitment*. Cambridge: Cambridge University Press, 1973.
Wittgenstein, Ludwig. *On Certainty*. Oxford: Basil Blackwell, 1969.
_____ *Philosophical Investigations*. New York: MacMillan, 1953.
_____ *Zettel*. Oxford: Basil Blackwell, 1967.
_____ *Philosophical Remarks*. New York: Barnes and Noble Books, 1975.
Wong, David B. "Three Kinds of Incommensurability," in *Relativism: Interpretation and Confrontation*, ed. Michael Krausz. Notre Dame: Notre Dame University Press, 1989.

The Southern Journal of Philosophy (1990) Vol. XXVIII, No. 4

CONVENTIONAL FOUNDATIONALISM AND THE ORIGIN OF NORMS

Ann E. Cudd
University of Kansas

Some theories of morality and language rely on an account of social conventions to give them a foundation in rationality.[1] They follow an argument schema which goes like this: 1. Moral (linguistic) norms are conventional; 2. the origin and maintenance of conventions can be explicated as rational solutions to coordination problems; 3. humans face critical coordination problems which are solved by morals (language); 4. therefore, the origin and maintenance of morality (language) is explicable as rational solutions to coordination problems. These theories thus rely on a foundational theory of convention, in which the foundation is instrumental rationality. I shall call the use of this argument schema *conventional foundationalism.* In this essay I argue that conventional foundationalism is fundamentally flawed, in particular, that premise 2 above fails.

David Lewis's account of the origin and maintenance of social conventions in *Convention: A Philosophical Study* is the *locus classicus* of the rational foundation argument for convention. Lewis sets out to refute the argument, made by Quine and others, that language cannot be conventional. The problem, simply stated, is that conventions seem to require agreements, and, so the argument goes, agreements presuppose an ability to communicate which is tantamount to having a language. So linguistic conventions can be constructed only on the foundations of language. Thus it seems that there is no good sense to the claim that language is ultimately conventional. Lewis attempts to refute this argument by showing how conventions can arise among rationally self-interested agents who have no prior conventions or norms. That is, he attempts to show that

Ann E. Cudd is an Assistant Professor of Philosophy at the University of Kansas. Her work is primarily in foundations of economics and rational choice theory, and their applications to political and moral philosophy. She has published articles in Public Affairs Quarterly, Theory and Decision, *and* The Journal of Philosophy.

rationality, with no previous communication or social interaction, is enough for agents to come to guide their behavior by conventions in situations which are called in game theory "coordination problems."

Lewis's theory of convention has broad appeal because it purports to provide a rational foundation for social interaction in the non-social. Thus a theory of morality for which conventional agreements are the primary kinds of moral norms can claim that morality is the result of the rational choice of the freest possible agent—one who need have no ties to, interest in, or even knowledge of, others before the agreement is made. In this essay I will argue that Lewis's theory of convention must make tacit appeals to pre-existing normative structures, (e.g., language), contrary to this strongest version of his claim. The implication of the necessity for such appeals is that convention cannot be the rational foundational basis for theories of language, morals, and agreements which Lewis and his foundationalist followers claim for it.

This conclusion is not completely original. Donald Davidson makes a similar claim when he writes, "philosophers who make convention a necessary element in language have the matter backwards. The truth is rather that language is a condition for having conventions."[2] The argument, which can be gleaned from points made in a number of essays,[3] is that in order to have conventions with other beings we have to be able to ascribe to them beliefs and intentions, and in order to ascribe beliefs and intentions to them we need to be, and know each other to be, language users. Thus language is a necessary condition for convention on Davidson's view. My argument differs significantly from Davidson's in three respects. First, the argument I offer directly attacks the internal consistency of Lewis's argument. I take it that Lewis offers the best attempt to provide a rational foundation for convention, and argue that the game theoretic argument fails to show that it is possible to reduce conventional interaction to actions of epistemically isolated rationally self-interested agents. Thus I question whether Lewis can provide game theoretic foundations of his thoery of convention, while Davidson's argument attempts to show that agents must have language to form conventions without addressing Lewis. Second, my argument attacks the use of conventions as a rational foundation for anything, not just language. So, for example, I will argue that conventions cannot provide rational foundations for morality. Third, while Davidson requires that language use and the mutual ascription of the ability to use

language is necessary, I do not claim that language is required for convention formation, and suggest that other normative structures may suffice as well. Although the arguments contain interesting disagreements, since our conclusion that convention cannot be foundational is substantially the same, I will not discuss Davidson's argument further. I will, rather, discuss the implications which this criticism has for foundationalist political and moral theories.

I. Lewis's foundational project and coordination problems

Lewis takes from Hume the claim that conventions result from a common sense of interest, so that promises are not needed for conventions to arise. Lewis aims to explicate Hume's argument by introducing the theory of games, in particular equilibria in coordination games, to show that conventions are the outcome of rational actions in such situations. He argues that although the theory of games can reveal the mechanisms of motivation and reason that guide conventional interaction, the theory is merely a scaffolding which can be kicked away once these mechanisms have been laid bare.

Game theory is a reductionist model, in which instrumental rationality and agents' beliefs and desires about the immediate situation are the only causal factors governing the agents' actions. Since it makes no presuppositions about the previous interactions of agents, if conventions can be shown to be the game theoretic solutions to some suitably identified interactions, then conventions will be grounded in instrumental rationality. The situations in which conventions arise are those representable as coordination games; Lewis calls such situations coordination problems. A coordination problem is a situation of strategic decision (i.e., the outcome for each depends on the strategies employed by all) for two or more agents in which there is a predominant coincidence of interests and at least two coordination equilibria. A coincidence of interests arises when all the agents are better off whenever any of them are, that is, when the prospects of each individual rise or fall with the prospects of all. For example, in driving it is best for each individual driver if everyone reaches their destinations safely. The more accidents there are the worse off each individual is because each has a greater chance of being an accident victim, and so each must drive more carefully and more defensively to reduce the risk. Thus actions which an individual or a group of individuals can take to lower their chances of an accident will benefit all drivers. By contrast, the prisoner's dilemma

is a situation in which one individual may gain at the expense of the other. Each prisoner does better if he confesses when the other does not, and the one who does not confess achieves his worst outcome. An equilibrium is a set of strategies (which for the purposes of this essay we can take to be identical to a set of actions), one for each agent, such that that strategy gives the agent the best outcome, given that the others will choose those strategies which achieve their best outcomes. A coordination equilibrium is a set of strategies such that no one would be better off if any one agent had acted otherwise.

As an example, imagine two drivers, Tim and Kim, who must decide on which side of the road to drive as they approach each other. Neither cares on which side they drive, so long as it is the same (relative) side as the other, and both gain the most from the interaction if they both drive on the same side. If we suppose that there are only two choices, namely drive constantly on the left or constantly on the right, then we can represent the game with the following payoff matrix.

		Kim: Left	Kim: Right
Tim	Left	1,1	-1,-1
	Right	-1,-1	1,1

There are two equilibria in this game, one where both drive on the left and one where both drive on the right, since in each case neither can do better given what the other is going to do. Both equilibria are also coordination equilibria since, given one equilibrium strategy, Tim cannot do better if either he does something else while Kim plays that equilibrium strategy or if she does something else while he plays that equilibrium strategy. So this game is a coordination game, since it contains at least two coordination equilibria. Since there are two equally good coordination equilibria, the solution, if there is to be a solution, will have to be in some sense arbitrary, and this is just the sense in which a convention is arbitrary.

In the rational choice model individual rational agents formulate strategy by recreating the thinking processes of the others. Rational individuals form estimates of what the others are thinking, and in this way anticipate and interpret each others' actions. Lewis describes the process of strategic thinking by rational agents when he writes:

> It is a process in which one person works out the consequences of his beliefs about the world—a world he believes to include other people who are working out the consequences of their beliefs, including the belief in other people who .... By our interaction in the world we acquire various high-order expectations that can serve us as premises. In our subsequent reasoning we are windowless monads doing our best to mirror each other, mirror each mirroring each other, and so on. (*Convention*, p. 32.)

Lewis's image of the rational individual is of one who uses her private beliefs about the world and others as inputs to manufacture a representation of the mutual beliefs and identify optimal strategies for action. Rational agents are thus essentially unconnected to others and to society, except in that they observe each other.

Lewis's project can be characterized now as an attempt to explicate convention in terms of the beliefs and desires of the parties to the convention. He characterizes coordination situations completely in terms of agent's desires, and their knowledge of each others' beliefs and desires determines whether and which coordination equilibrium becomes a convention. Recall that equilibria are sets of strategies that are, when deployed by all the agents in combination, optimal for each agent. So equilibria are appealing solutions to coordination problems, since they remain optimal no matter how many times rational agents "mirror each other, mirror each mirroring each other, and so on." But in coordination situations it seems that there is room for vacillation on the part of the agents because there are multiple equilibria. In the driving game Tim and Kim have two equally good coordination equilibria. Although they can find the equilibria alone, they also need a way to find the right one to aim for in this particular interaction. Lewis's task is to show how a particular coordination equilibrium becomes the stable, consistent choice of agents without a previous history of interaction.

According to Lewis there are three ways that agents come to conventions: by agreement, through recognition of some salient action which all expect the others to perform, and through precedence. Agreement is by far the most obvious and convenient means of arriving at a convention; if Tim and Kim happen to meet to discuss their impending interaction they can simply agree on one side or the other, and this will be enough to give them a convention for coordinating their actions in the future. Precedence may also do this for them, so that if they drove on the right side last time they passed they may continue to do so, provided they think that each reaches the same conclusions about what precedent holds. If

neither of these is available they might think that one side is salient because of the preponderance of right-handed drivers, or the fact the their steering wheels are on the left side of the car, or whatever, but again they must trust that each finds the same actions salient, and acts accordingly.

To list these three mechanisms for finding conventions is not really to explain much, however. For we have not yet said much about what it is to be a rational agent, or about why these mechanisms give rational agents good reasons to rely on their conclusions about what they ought to do. How is it that these give "windowless monads" enough information for them to solve coordination problems? Rational choice theory supposes that agents are rational in the minimal sense that they perform that action which maximizes their expected utilities[4] given their beliefs. Included among the beliefs that an agent has are her beliefs about others' beliefs. In order to trust that the others will do their part in coming to the coordination equilibrium, one has to expect that the others expect that each will do her part, which requires that each expect that each expect that each will do her part, and so on. That is, the agents need to have a (potentially infinite) set of nested mutual expectations about each other that the others will do their part.

To see that all of these expectations are really necessary, imagine that Tim and Kim have passed each other once before and they happen to have driven on the right side at the time, which was also on a day with an even-numbered date. Tim might suppose that the precedent is to drive on the right, but Kim might suppose that the precedent is to drive on the right on even days, and on the left on odd days. In that case they will not successfully coordinate on odd days. Now suppose that they both decide on the same precedent, but they are unsure whether the other has decided on the same precedent. Then they are no closer to finding a solution, since Kim will drive on the right(left) if she knows that that is the precedent, unless she suspects that Tim thinks that driving on the left(right) is the precedent, and likewise for Tim. They face a new coordination problem—that of choosing the precedent. So as long as they are unsure of the other's (first order) expectations, they are unsure of what to do. To illustrate one more order, suppose that Tim and Kim have the required first order expectation, but lack any second order expectation about the other, that is they do not know what to expect that the other expects. Then they do not know whether the other one is in the situation described just previously, i.e., the one in which they were missing the first order expectation, and so

could not solve the problem. Thus they expect that the other may not solve the problem correctly, which means that they must rethink the solution, and now the best they can do is guess. I leave it to the reader to check the higher order expectations to see that they are crucial to the solution of this problem. Just as they need to have the right mutual expectations abut what coordination equilibrium to choose, they also need to have expectations and mutual expectations about each other's rationality.

Lewis claims, then, that a characteristic feature of all conventions is that they are (by and large) commonly known, where by common knowledge he means that a fact f is common knowledge among a population P if and only if all the members of P know f and they all know that they all know f and they all know that they all know that they all know f, and so on. Conventions coordinate actions of people who are party to the conventions by providing them with the mutual expectations that they need to find equilibria in coordination problems. Agreement, salience, and precedence aid agents in picking out which equilibrium, among otherwise equally good choices, on which to convene. However, for Lewis to claim that conventions are strictly speaking *solutions* to coordination games, then these mutual expectations must exist for each order, that is to say, the choice called for by the agreement, precedent, or salience, as well as the rationality of the players, must be common knowledge.

II. Convention

Common knowledge of many facts is thus necessary for conventions to arise in a population. In order to *ground* convention in the theory of games, that is, in order to derive the convention without presupposing some other conventions or norms, Lewis needs to argue that common knowledge can be generated among a population of rational individuals who share no norms or knowledge of each others' norms. Additionally, he needs to show that the generation of common knowledge is rational, i.e., utility maximizing to individuals. On Lewis's view we need to find a *basis* for common knowledge in a population P, in some state of affairs, A. The basis is a set of nested reasons to believe, which, when supplemented by mutual ascriptions of rationality, gives us the required nested levels of mutual knowledge. On Lewis's view A must meet the following conditions:

(1) The individuals in P have reason to believe that A holds.

(2) A indicates to those individuals that they each have reason to believe that A holds.
(3) A indicates to each _____.

The blank space in (3) stands for the items of common knowledge (in this case the contents of the conventions). Lewis defines the relation of indicating as:

(D1) A indicates to individual x that _____ if and only if, if x had reason to believe that A held, x would thereby have reason to believe that _____.

What A indicates to the individuals depends on their background knowledge, shared norms and conventions of induction and inference, etc. In other words, in order for a state of affairs to be a basis for common knowledge, the individuals must already have some common knowledge about each other. Conditions (1)-(3) and (D1) generate an infinite set of nested "reasons to believe":

(1′) Each has reason to believe _____.
(2′) Each has reason to believe that each has reason to believe _____.
(3′) Each has reason to believe that each has reason to believe that each has reason to believe _____.

These nested inferences about reasons to believe become beliefs if there is a sufficient expectation of mutual rationality. Suppose that each is rational. Now we need the inference rule that if x is rational, then x believes whatever she has reason to believe, and we need this to be common knowledge. Then we have

(1″) Each believes that _____.
(2″) Each believes that each has reason to believe _____.
(3″) Each believes that each has reason to believe that each has reason to believe that _____.

And so on. To further convert reasons to believe into beliefs we need to have more levels of belief of rationality. We would then end up with:

(1‴) Each believes that each believes that _____.
(2‴) Each believes that each believes that each has reason to believe _____.
(3‴) Each believes that each believes that each has reason to believe that each has reason to believe that _____.

Lewis recognizes that the generation of mutual beliefs based on A can go only so far as the ancillary mutual expectations about rationality and inference rules. Ascribing infinitely many levels of these further mutual beliefs (about rationality and the inference rules) is only a problem if they cannot be implicitly held, i.e., if they must be actual beliefs held by agents, not just implications of beliefs. The point of positing the basis for common knowledge is that it implicitly represents the levels of reasons to believe. Lewis posits another basis for the common knowledge of mutual rationality. He supposes that the individuals of P could have a basis that gives them reason to believe that each is rational (e.g., from observing their behavior) and that each believes that each is rational and that each believes that each believes . . . . But this basis can only generate reasons to believe, and hence requires antecedent beliefs about the individuals' rationality and their mutual ascriptions of rationality, and once again the inference rules, to generate beliefs. Thus it seems that we find ourselves in a regress if we try to discover rational processes of common knowledge formation.

Is the regress vicious? Lewis denies that it is. He writes:

> A basis for common knowledge generates higher-order expectations with the aid of pre-existing higher-order expectations of rationality. Can these themselves be generated by some basis for common knowledge? Yes, because all the higher-order expectations of rationality needed to generate an nth-order expectation are themselves of less than nth-order.[5]

Lewis's claim is that a basis for common knowledge bootstraps the necessary expectations of rationality. The bootstrapping operation on rationality works as follows.[6] Assume that A is a basis for common knowledge, which means that: (1) each agent has reason to believe that A, (2) A indicates that all agents have reason to believe that A, and (3) A indicates all agents are rational. Assume also, of course, that the agents are rational. Now take as one inference rule that if agents are rational and they have reason to believe that q, then they believe that q, and another inference rule that if p strictly implies q then if an agent i has reason to believe p then (strictly) she has reason to believe that q. It is now straightforward to show that the bootstrapping can be carried out. (See appendix for the proof.)

But let us look closer at the notion of strict implication in this definition of indicating. Recall that the definition is

(D1) A indicates to individual x that _____ if and only if, if x had reason to believe that A held, x would thereby have reason to believe that _____.

Lewis (1973) defines strict implication as follows: p strictly implies q iff $\Box (p \Rightarrow q)$.[7] A sentence $\Box$ p is true at a world i if and only if it is true at all worlds accessible to i. The strictness of an implication varies with the accessibility relation: the more worlds that the accessibility relation includes, the stricter is the implication.

Although he does not say how strict the implication has to be in his definition of indicating, the choice is crucial if we want conventions to have a purely rational foundation; the bootstrapping cannot be carried out if it is the weakest implication, where only this world is accessible, i.e., material implication, since one would then need the rule 'if p *materially* implies q then one's having reason to believe p *materially* implies that one has reason to believe q', which is clearly fallacious. (To see that it is fallacious just take p to be 'grass is green' and q to be 'snow is white'.) The inference rule needs a stronger sense of implication in order to have any plausibility. The stricter the implication is, the stricter the corresponding implication may be in the inference rule. However, the stricter the implication, the more work the basis for common knowledge has to do in its indicating role. So what he needs is a sense of strict implication which is strong enough for the inference rule to be plausible, and yet not so strict that the definition of indicating is implausible.

I claim that there is *no* sense of strict implication which meets Lewis's needs. In order for his agents to have the logical wherewithal to make the inferences that windowless monads need to make, Lewis must endow them with knowledge of all logically valid inference rules. This knowledge must be theirs by virtue of being rational, since they have no *prior* basis for common knowledge of inference rules, on pain of regress. So if they know of each other that each is rational, they know that all know these inference rules.[8] More than that, these inference rules must be common knowledge, so that all the higher order inferences can be made. Such a conception of knowledge is represented in Hintikka's axioms for knowledge,[9] in which the knowledge operator, 'K', which may be translated as 'knows that,' is an S5 modal operator. In such a logic it is an axiom that all logical truths are known by all agents, and since that is a logical truth, it is known by all agents that all agents know all logical truths, and so on.

Furthermore, the strict implication of the indicating function must be something very like logical implication—it must be unthinkable that anything other than what is indicated by A to one person is indicated to another; they

must believe that it is a logical implication. Otherwise there is room for doubt about what the basis indicates, and this can collapse the whole web of mutual beliefs about what A indicates. Perhaps we should call this "strict epistemological implication." If the definition is read as using strict epistemological implication, then if A indicates _____ to one agent, she can be sure that A indicates _____ to all agents and that that is known for any finite order of mutual knowledge (i.e., 'I know that you know that I know . . . '), since all logical truths are mutual knowledge at any finite order on the strong conception of rational agency which Lewis utilizes.[10]

The stronger version of the definition of indicating assumes more on the part of agents than we ought to assume if we want to bootstrap common knowledge from a non-social origin, however. In particular, it implies common knowledge of what bases indicate. There are two possible grounds for choosing strict epistemological implication: empirical and normative.[11] One might claim that it is an empirical fact that bases indicate the same thing to everyone. But this is not plausible, since we are all well acquainted with misunderstandings, intended and unintended, which even as solid a base as language can give rise to. Metaphor, puns, poetry, inside jokes all have in common the ability to indicate many different things to different people. And since we all know this, we are usually on the lookout for it. One might make a normative claim to justify the strong interpretation of the indicating function: having a reason to believe something means that it should also be a reason for others to believe it. What force does the "should" have here? Again, it is not plausible that it is a disguised descriptive claim, so it must be a reference to some rules for giving reasons. That is, the claim is that one cannot be said to have a reason to believe p unless one can then give others reason to believe p. But then one has to know what counts as reasons for others. On pain of regress this knowledge is either empirically given or it requires a preexisting normative structure, (such as a shared language, or shared scientific or religious beliefs and standards of evidence). And once again it is not plausible to suppose that it is empirically given with the certainty and universality required. Norm making, though, is an essentially social enterprise. To understand the claim normatively, then, assumes a social dimension of rationality to which a foundationalist is not entitled. We have conventions and linguistic norms for discerning linguistic signification and implications, but the foundationalist cannot help himself to

such things if he wants to explain the conventional nature of language by reducing it to game situations in which common knowledge emerges.

One might object that Lewis needs only one basis to bootstrap the common knowledge starting point, and then common knowledge production is up and running on its own. Contextual features of situations help us to decide what is meant in these linguistic cases, and the context could provide further bases for common knowledge. This can happen only if these features themselves indicate their significance to all rational agents. We should recognize this as entailing the myth of the given:[12] the idea that raw sense data are given to rational beings, without any taking or interpretation by the beings themselves. This assumption would entail that our interpretations are forced, by the data itself, to be identical. But if that were the case we would all conform because of our similar reactions to stimulus; common knowledge would drop out of the explanation entirely. Furthermore, he might be taken to give a rational choice *justification* of common knowledge, but the justification then rests on a strong empirical assumption about our physiological reactions to stimuli. The assumption, if true, would *explain* the required common knowledge, but could not serve to justify it as a conclusion at which rational agents arrive. So it seems that we can either believe in the myth or we can, as another philosopher has suggested in a similar context, just stop tugging at our bootstraps altogether.[13]

To summarize the criticism of Lewis's theory of convention: his theory was designed to explain conventions as rational choices of isolated agents, specifically, as equilibria in coordination games in which there are multiple equilibria. If successful, then conventions, which seem to involve agreements (which presuppose social interaction) would be explicable by a plausibly non-social feature of humans, namely our instrumental rationality; a seemingly highly social aspect of our behavior is explained as the outcome of rationally self-interested behavior. However, in order to make this explanation it is necessary to posit common knowledge among the agents party to the convention. Lewis then tries to construct common knowledge purely rationally, in order to avoid bringing in any new conventionality through the back door. I have tried to show that his construction still depends on a prior assumption of common knowledge. Thus I conclude that on Lewis's theory common knowledge, and *a fortiori* convention, arises only in a context in which there is common

knowledge of what is indicated by various bases, and this assumes the prior existence of norms.

III. An illustrative example

What does this mean for the uses to which Lewis's foundational theory of convention has been put, i.e., the conventional foundationalism I spoke of at the beginning? I shall try to answer this by way of an illustrative example.

The example compares two different explanations of the origin and maintenance of conventions of war. In *The Economic Theory of Social Institutions*, Andrew Schotter explains the reluctance of nations to wage all-out war as conventional sanctioning rules which are, in effect, cooperative solutions to indefinitely repeated prisoner's dilemmas. He argues that nations use less destructive weapons systems in wars whenever they thereby maximize their long run payoffs, given the payoffs of the current war (or battle), the expectations that the others will follow a similar cooperative strategy, and the payoff of all-out war on both sides. A simple model of this situation is as a prisoner's dilemma (PD) in which the two strategy choices of the players are to either wage 'limited' or 'all-out' war. The payoffs are given in the matrix below.

		Nation B: limited	Nation B: all-out
Nation A	limited	5,5	-1,10
	all-out	10,-1	0,0

Since PD cooperative outcomes (in this example, both wage limited war) are not equilibria, i.e., not best responses given what the other is doing, they require some sort of policing mechanism if they are to be stable. That is, there must be some sanctions which change the game from a PD to some game in which the cooperative outcome is a stable equilibrium. Thus Schotter writes that they "must involve a sanctioning rule that specifies what reaction the other n-1 players will have when any given player deviates from the behavior specified by the institution."[14]

By modeling the game as an *iterated* PD (IPD) we change the game to one in which such rules can arise as part of the strategies of the game. Schotter models the war supergame

as an indefinitely repeated game, that is, the PD situation is repeated each time with a probability between zero and one, exclusively. It has been shown in such games that, given a small enough probability that the current iteration is the last one, there will be cooperative equilibria. There are many possible equilibria, some of which are equally good. Hence this is a coordination problem, and the resulting solution is a convention. For example the so-called tit-for-tat strategy is an equilibrium strategy for a large class of these IPD games. The tit-for-tat strategy tells the agent to begin by waging a limited war, then in the next war wage limited war if the other nation was similarly restrained in the previous war, and wage all-out war if the other did in the previous war. This strategy is an example of what Schotter means by a sanctioning rule, since waging all-out war is a way to sanction others who do not reciprocate one's limited war.

Schotter wants to explain how these conventional sanctioning rules evolve by showing that they constitute coordination equilibria. There is an ambiguity in 'evolve'; it may mean either the process of evolution or the end product, that is, either why the rules are thought of or why they are not rejected. Schotter is trying to explain the former. Thus he needs to show that because the rules constitute equilibria they provide an incentive for the players to play those rules. However, as we have seen, in order for any particular set of strategies to be in equilibrium, they, or at least the available strategies, must be common knowledge. But it seems that the best explanation of such common knowledge (if, indeed, this common knowledge ever arises) would be that there is a convention about what strategies are normally played. In that case the convention explains the equilibrium, and not the reverse.

What I am criticizing here is the explanation of the origin of the convention, I am not doubting that such conventions exist. The convention is not rationally arrived at, though it may be rationally maintained. The crucial mistake of both Lewis and Schotter is to conflate the explanation of the maintenance of norms with the explanation of their origin. There are other possible explanations of the origin of convention. In his book *The Evolution of Cooperation*, Robert Axelrod recounts details of trench warfare in World War I between French and German troops. In this highly non-cooperative situation a startling example of cooperation in a PD-like situation arose: there came to be conventions concerning when one would shoot at the other trenches, and when artillery shells were to be launched. The soldiers were

tacitly cooperating in order to enhance survival rates to the effect of deescalating the war. This is an example of what Schotter means by a convention of war. But the information they had about each other was minimal, at least until the cooperation had already begun, when they could infer that the others wanted basically the same thing as themselves. Axelrod explains this cooperation in terms of trying out a tit-for-tat strategy, which resulted in what could be described, *post hoc*, as an equilibrium solution. In this case equilibrium seems to have actually resulted from repeated play by myopic, but cautious, players. That is, the soldiers tried less aggressive strategies, and as long as that behavior was returned, they continued. Their retaliations were also not a matter of rational foresight aimed at preserving an equilibrium, but rather a matter of a code of honor among their own. Axelrod writes: "a powerful ethic of revenge was evoked. This ethic was not just a question of calmly following a strategy based on reciprocity. It was also a question of doing what seemed moral and proper to fulfill one's obligation to a fallen comrade."[15] After some time the stability of this behavior was noticed and only then could it be said that the equilibrium was a motivation in itself.

The difference in the two explanations is that on Axelrod's theory the payoff functions and punishment strategies need not be common knowledge, while on Schotter's theory the analysis that the game theorist does is itself a motivation for the players. The difference between the two theories might be made clearer with an analogy. On one theory evolution works by a mechanism in nature which optimally adapts, or molds, the species to its environment. On the other theory evolution occurs as a result of the differential reproductive success of individuals with certain characteristics, which *post hoc* we call adaptive; the characteristics are not optimal globally, but only locally.[16] The difference is subtle, but it is the difference between an illegitimate teleological explanation in the first place, and a legitimate causal explanation in the second. In the disagreement between Schotter and Axelrod we are looking at two species of intentional explanations, one which requires global optimization while the other requires only locally optimal behavior. Axelrod's is the latter, and I believe, clearly the better in any situation in which there is no complex normative structure, or in which that structure has broken down. The soldiers in Axelrod's example may well be quite rational, but the norms to which they adhere evolve unconsciously, prior to their rationality, as one should expect

in an environment in which there is no prior reason to think that trust or even reliable communication is available.

Schotter views the information situation as being much richer than the game theoretical analysis allows. He writes:

> In the final analysis, when countries fight wars, they have more information at their disposal upon which to make strategic choices than merely the strategic capabilities of the conflicting parties and their preferences. They have, in addition, knowledge of a whole set of tacitly agreed to rules of conduct or conventions of war to which they adhere.[17]

Without the assumption of these conventions of war the strict common knowledge requirements of Schotter's theory are lacking. Thus his game theoretic explanation can do only part of the explanatory work; he does not explain the origin of the conventions, but only that once there are such conventions it is in the best interest of the participants to adhere to them. There is, indeed, much more contextual information than the game theoretic analysis uses, as he points out in this quote. But this contextual information is itself the conventional knowledge that needs to be explained, and in order to show how these conventions arise one needs something other than this model of agents as rationally self-interested isolated decisionmakers.

IV. Conclusion

The lesson for rationalist foundational moral and political theories from this discussion is that if conventional agreement provides the foundation then the theory faces the following dilemma. Either the conventions are themselves founded on some pre-existing normative structure which erects the necessary common knowledge, or the theory must presuppose givenness. If the theory presupposes a normative structure, then it is not foundational, but perhaps that is preferable to presupposing givenness. Rousseau, in his *Discourse on the Origin of Inequality*, provides us with a naturalistic explanation of normative structures which do not ultimately rely on rationality; early humans learned to interact through trial and error on his account, much like Axelrod's soldiers. Hume, we recall, presupposes a natural moral sentiment which would explain the similarity of reactions of humans to similar social situations; he gives us a physiological basis for givenness. Their theories may be objectionable on other grounds, but they have avoided this dilemma by grasping one or the other of its horns.

The dilemma raises serious questions about projects like that of Hobbes, or its contemporary development in David

Gauthier's *Morals by Agreement*, if one takes them to attempt to ground morality in instrumental rationality, as Lewis tried to ground conventional behavior in rationality. What I have shown here is that there is a problem with the foundational rational choice account of convention, and these theories are not conventional accounts of morality. However, they rely on similar rational choice account of norms, which may turn out to be equally unjustifiable.

It is clear that there are some conventions which solve coordination problems. This essay questions whether these solutions can be shown to arise among isolated rational agents. I have argued that common knowledge is necessary for an explanation which requires a procedure for the agents to work out the game in the head, or for an explanation based purely on selfish interests and strategic rationality. There are, of course, other explanations for the solutions, as Axelrod's myopic players demonstrate. In general I would suggest that social and political conventions arise by extension from other norms which provide all the complex mutual knowledge required. I have simply tried to show that conventions cannot be founded on the rational choices of windowless monads, and that this implies that one cannot rely on conventional behavior to show how the social can be founded on the non-social alone.[18]

Appendix: Lewis's derivation of beliefs from the reasons to believe provided by a basis for common knowledge.

Let H(i,q) stand for "i has reason to believe that q," let R(i) stand for "i is rational," and let B(i,q) stand for "i believes that q." Let '⇒' be the conditional, as before, and let '→' stand for strict implication, as discussed in the text.

Inference rules:

(r1) (Ri & H(i,q)) → B(i,q)
(*) if p → q, then H(i,p) → H(i,q)
(**) if (p & q) → s, then (H(i,p) & H(i,q)) → H(i,s)
(***) if (p & q & r) → s, then (H(i,p) & H(i,q) & H(i,r)) → H(i,s)
.
.
.

Assumptions:

(1) (i) H(i,A)
(2) A indicates (i) H(i,A)

(3) A indicates (i) Ri
(4) (i) Ri

(1)-(3) are just what is meant by the claim that A is a basis for common knowledge. Lewis translates (2) and (3) according to the definition of indicating as follows:

(D1) A indicates to x that _____ iff H(x,A) → H(x,_____). Replacing _____ by the content of (2) and (3), and x by j we get,

(2) H(j,A) → H(j,(i) H(i,A))
(3) H(j,A) → H(j,(i) Ri)

To bootstrap the reasons to believe that the others are rational we do the following derivation.

(2.1) H(j,A) → H(j,H(i,A))	2
(2.2) H(i,(H(j,A)) → H(i,H(j,H(i,A)))	2.1,(*)
(2.3) H(j,H(i,H(j,A))) → H(j,H(i,H(j,H(i,A))))	2.2,(*)
.	
.	
.	
(1.1) H(j,A)	1, repeated
(1.2) H(j,H(i,A))	1.1,2.1
(1.3) H(j,H(i,H(j,A)))	1.2,2.2
(1.4) H(j,H(i,H(j,H(i,A))))	1.3,2.3
.	
.	
.	
(4.1) Rj	4, repeated
(4.2) H(j,Ri)	1.1,3.1
(4.3) H(j,H(i,Rj))	1.2,3.2
(4.4) H(j,H(i,H(j,Ri)))	1.3,3.3
.	
.	
.	

Now to get from having reasons to believe to third level beliefs in each others' rationality:

(5.1) [H(j,Ri) & Rj] → B(j,Ri)	(r1)
(6.1) (H(i,H(j,Ri)) & H(i,Rj)) → H(i,B(j,Ri))	5.1,(**)
(5.2) [H(i,B(j,Ri)) & Ri] → B(i,B(j,Ri))	6.1,(r1)
(6.2) [H(j,H(i,B(j,Ri)) & Ri)] → H(j,B(i,B(j,Ri)))	5.2,(***)
(5.3) [H(j,B(i,B(j,Ri))) & Rj → B(j,B(i,B(j,Ri)))	6.2,(r1)
.	
.	

.	
(7.1) B(j,Ri)	4.2,4.1,5.1
(7.2) B(i,B(j,Ri))	4.1-4.3,5.2
(7.3) B(j,B(i,B(j,Ri)))	4.1-4.4,5.3
.	
.	
.	

So it appears that Lewis can bootstrap common knowledge from the assumptions that the players are rational. Note that inference rule (*) is the one at issue in the discussion of the strict conditional, '→', in the text. In order for the derivation to go through, the conditional in the definition of 'indicating' must be translated as a strict implication, '→', rather than the sentential connective, '⇒'. The bootstrapping cannot be carried out if assumption (2) is: H(j,A) ⇒ H(j,H(i,A)), unless we use the clearly fallacious rule: (p⇒ q) ⇒ [H(i,p) ⇒ H(i,q)].

NOTES

[1] Hampton (1986), Ullman-Margalit (1977), Schotter (1981), Gauthier (1979), and Sugden (1986), are some examples of works that lean heavily on a foundational account of convention as a foundation for moral and social theories, as well as the account of language in Lewis (1969).

[2] Davidson (1984), p. 280.

[3] See Davidson (1984), especially the essays: "Thought and Talk," "On the Very Idea of a Conceptual Scheme," and "Communication and Convention."

[4] This is more general than the theory that Lewis uses, since he never deals with mixed strategies or uncertainties, but it does no harm to suppose this conception of utility maximization.

[5] *Ibid.*, p. 57.

[6] The clarification was made in private correspondence.

[7] Lewis (1973), p. 4. See sections 1.2-1.3 for the account of strict implication to which the discussion in the text and the appendix refers.

[8] This argument is discussed extensively in Cudd (1988). The point is that in order for a coordination game to have an equilibrium, the game must be common knowledge, as well as the rationality of the players and their payoffs. For the rational foundationalist to make the argument that rationality can get the agents to a convention by rationality alone, they need common knowledge of these things as well.

[9] See Hintikka (1962).

[10] In fact, we might be able to replace 'mutual knowledge of any finite order' by 'common knowledge' if a slightly stronger set of axioms is supposed. (See Cudd, 1988.)

[11] Lewis does not choose either of these since he does not see that the indicating function will need to indicate the same thing to everyone. He writes: "What A indicates to x will depend on x's inductive standards and background information" (p. 52). The following two paragraphs in the text represent what the conventional foundationalist might say to save the rational foundation of the theory.

[12] Sellars (1963).

[14] Schotter (1981), p. 41.
[15] Axelrod (1984), p. 85.
[16] See Elster (1984).
[17] Schotter (1981), p. 41.
[18] I would like to thank David Gauthier, Tamara Horowitz, David Lewis, Arthur Ripstein, members of the University of Kansas Philosophy Faculty Colloquium, and an anonymous referee of the *Southern Journal of Philosophy* for helpful comments on earlier drafts of this essay.

REFERENCES

Axelrod, R., 1984, *The Evolution of Cooperation*, New York: Basic Books.
Cudd, A. E., 1988, "Common Knowledge and the Theory of Interaction," University of Pittsburgh Ph.D. dissertation.
Davidson, D., 1984, *Inquiries into Truth and Interpretation*, Oxford: Clarendon Press, pp. 265-280.
Elster, J., 1984, *Ulysses and the Sirens*, New York: Cambridge University Press.
Gauthier, D. P., 1979, "David Hume: Contractarian," *The Philosophical Review*, vol. 88, pp. 3-38.
Hampton, J., 1986, *Hobbes and the Social Contract Tradition*, New York: Cambridge University Press.
Hintikka, J., 1962, *Knowledge and Belief*, Ithaca: Cornell University Press.
Lewis, D. K., 1969, *Convention, A Philosophical Study*, Cambridge: Harvard University Press.
....., 1973, *Counterfactuals*, Oxford: Basil Blackwell.
Quine, W. V. O., 1980, "Two Dogmas of Empiricism," in *From a Logical Point of View*, Cambridge: Harvard University Press.
Schotter, A., 1981, *The Economic Theory of Social Institutions*, New York: Cambridge University Press.
Ullman-Margalit, E., 1977, *The Emergence of Norms*, Oxford: Clarendon Press.
Sellars, W., 1963, "Empiricism and The Philosophy of Mind," in *Science, Perception and Reality*, New York: Humanities Press.

The Southern Journal of Philosophy (1990) Vol. XXVIII, No. 4

LANGUAGE AND POLITICAL AGENCY: DERRIDA, MARX, AND BAKHTIN

Fred Evans
Iowa State University

In "Structure, Sign and Play in the Discourse of the Human Sciences" (1978), Derrida distinguishes between two types of interpretation, one which seeks the origin of meaning and truth outside the play of signs in which such values must be expressed or indicated, and another that eschews origins, affirms the play of signs, and strives only to effect the passage from one signifier to another without end.[1] To the degree that these two types of interpretations concern themselves with dislodging an ideology or a political order, they contain specific views of the relation between language and political agency. In the type of interpretation that seeks origins, language is usually depicted as a tool through which political agents (individuals or classes) can express what they mean, pronounce the truth, and fulfill their destiny. In the second type of interpretation, the self and meaning are so many effects of the play of signs, so that agency is more properly attributed to language than to persons. Derrida particularly associates this type of interpretation with "Nietzschean *affirmation*, that is, the joyous affirmation of the play of the world and of the innocence of becoming, the affirmation of a world of signs without fault, without truth, and without origin, which is offered to an active interpretation" (1978, p. 292).

As Derrida has indicated in numerous articles, his own aim is to "deconstruct" the first type of interpretation, that is, to overturn ("to bring low what was once high," 1981, p. 42) the dominance of the Western metaphysical tradition, of "phonocentricism," "logocentricism," and "phallocentricism." The first phase of this deconstruction consists in showing that any origin or teleological end is inseparable from the play of signs, from textuality, and that it is therefore not

---

Fred Evans received his Ph.D. in Philosophy from The State University of New York at Stony Brook and is an Assistant Professor of Philosophy at Iowa State University. He has published journal articles, book chapters and book reviews in the areas of Continental Philosophy, Philosophy of Psychology, and Philosophy of Technology.