History of European Ideas 37 (2011) 325–335

Contents lists available at ScienceDirect

History of European Ideas

journal homepage: www.elsevier.com/locate/histeuroideas

Normativity and Instrumentalism in David Lewis' *Convention*

S.M. Amadae *

Department of Political Science, The Ohio State University, 2140 Derby Hall, 154 N. Oval Mall, Columbus, OH 43210, USA

ARTICLE INFO

Article history:
Available online 17 December 2010

ABSTRACT

David Lewis presented *Convention* as an alternative to the conventionalism characteristic of early-twentieth-century analytic philosophy. Rudolf Carnap is well known for suggesting the arbitrariness of any particular linguistic convention for engaging in scientific inquiry. Analytic truths are self-consistent, and are not checked against empirical facts to ascertain their veracity. In keeping with the logical positivists before him, Lewis concludes that linguistic communication is conventional. However, despite his firm allegiance to conventions underlying not just languages but also social customs, he pioneered the view that convening need not require any active agreement to participate. Lewis proposed that conventions arise from "an exchange of manifestations of a propensity to conform to a regularity" (87–8).

In reasserting the conventional quality of languages and other practices resting on mutual expectations, Lewis comfortably works within the analytic tradition. Yet he also deviates from his predecessors because his conventionalist approach is comprehensively grounded in instrumentalism. Lewis adopts an extension of David Hume's desire-belief psychology articulated in rational choice theory. He develops his philosophy of convention relying on the highly formal mid-twentieth-century expected utility and game theories. This attempt to account for language and social customs wholly in terms of instrumental rationality has the implication of reducing normativity to preference satisfaction. Lewis' approach continues in the trend of undermining normative political philosophy because institutions and practices arise spontaneously, without the deliberate involvement of agents. Perhaps Lewis' *Convention* is best seen as a resurgent form of analytic philosophy, characterized by "a style of argument, hostility to [ambitious] metaphysics, focus on language, and the dominance of logic and formalization" that solves the dilemma of "combining the analytic inheritance...with normative concerns" by reducing normativity to individuals' preference fulfillment consistent with the axioms of rational choice.

Normativity in David Lewis' *Convention*

> In our subsequent reasoning we are windowless monads doing our best to mirror each other, mirror each other mirroring each other, and so on. – David Lewis[1]

David Lewis' 1969 *Convention* became noted as a brilliant contribution to theorizing about social customs, norms, and institutions and presented a novel approach to understanding language. Two points about this text are salient. First, it wholly adopted the new theoretical paradigm of game theory to construct its central argument that social conventions can be explained in terms of the underlying structure of individuals' preferences over outcomes. Second, this analysis implies that norms are strictly a product of instrumental rationality. Lewis' elegant monograph suggests that the social contract and other types of organized activity arise and persist independently from practices of agreement, promising, or tacit consent. Although Lewis claims that his insight follows from David Hume's conventional account of government, and that his use of game theory offers no new point of departure from Hume's original study, one goal here is to investigate the extent of the novelty of Lewis' approach. I am specifically interested in understanding the fate of the concept of normativity that follows from accepting Lewis' approach to convention.

Normativity, or the felt obligation that compels individuals to act in accordance to norms, is central to social organization. The reduction of normativity to instrumental rationality, and the resulting analysis of social institutions, is a major theme in late-twentieth-century social science developed by the epistemic community accepting rational choice theory. As this trajectory of research developed into the twenty-first century, it has articulated a view of society that is structured by norms which function as equilibrium solutions to games in which each individual pursues self-interest. The idea that agents could

* Corresponding author. Tel.: +1 614 292 9986.
E-mail address: amadae.1@osu.edu.

[1] David Lewis, *Convention*, Blackwell (Oxford [1969], 2002), 32.

0191-6599/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.histeuroideas.2010.10.011

deliberately work together to achieve common goals is replaced by the perspective that sociability and agreement are epiphenomenal to a base of individualistic and possibly unthinking maximization of expected utility.[2]

Game theory and expected utility theory are used to model and explain behavior, as well as to characterize the prescriptive norms specifying ideal purposive agency.[3] Moreover, as a proxy for instrumental rationality, game theory is postulated to be value free. Yet, as I hope to make clear in articulating Lewis' contribution to understanding norms and language, this supposedly neutral science of decision has the implication of dismissing alternative species of action. Agents obeying the axioms of rational choice do not voluntarily abide by rules or norms that they deem appropriate if the opportunity arises to cut a corner directly in the service of self-interest.[4] The traditional view of normativity, which accepts that agents may conform to rule-like conventions out of a motive independent from maximizing gain on a momentary basis, is anathema to both the prescriptive and explanatory branches of game theory.[5] Lewis' *Convention* extended the exhaustive instrumentalist view into the domain of social customs and the interstices of linguistic performance.

Lewis' *Convention* bears a closer relationship to his analytic predecessors than other expressions of post-positivist philosophy discussed in this issue. Embracing the thinnest metaphysics that exceeds strict verificationism, he is best regarded as a paradigmatic analytic philosopher. He accepts the meaningfulness of discussing possible worlds which cannot be proven to exist. He counters the dilemma of encompassing normativity in his analysis by reducing it to instrumentalism. He maintains the traditional earmarks of analytic philosophy: a particular argument style, a thoroughgoing naturalism, focus on language, and a kinship with logic and formalization. Rational choice theory is overtly formal, although Lewis' embrace of it is relatively discursive. Possibly of the various successors to the analytic tradition, Lewis remains one of its truer proponents. Expected utility theory and game theory owe their credibility entirely to their analytic structure. Lewis' new philosophic contribution is to join a predilection for analytic argument with instrumentalist explanations to provide an account of social conventions.

Background to Lewis' *Convention*

David Lewis, who was a doctoral student of W.V.O. Quine, drew on Thomas Schelling's game theoretic treatment of conflict. Game theory, first articulated in John von Neumann and Oskar Morgenstern's *Theory of Games and Economic Behavior* (1944), represented a purely mathematical theory of strategic rationality. The theory is comprehensive in its demands upon reasoners and its promise to identify the solutions to strategic interaction:

> We described…what we expect a solution – i.e. a characterization of "rational behavior" – to consist of. This amounted to a complete set of rules of behavior in all conceivable situations. This holds equivalently for a social economy and for games.[6]

Game theory came to be applied to a wide variety of social circumstances ranging from pure conflict in which one person's gain is the other's loss, to pure coordination in which if one benefits, so does another. Its initial promise had been to ground the science of warfare, at least in its development at the military think tank, the RAND Corporation.[7] Its earliest and possibly greatest policy triumph was in justifying Secretary of Defense Robert S. McNamara's nuclear strategy of Assured Destruction. The intellectual basis for this national security policy was articulated by Schelling, who served in the US Department of Defense under McNamara in the early-1960s.[8]

David Lewis was directly inspired by Schelling's *Strategy of Conflict* (1960) as he developed a game theoretic approach to social conventions, viewing them as coordination games. Much of the initial excitement about game theory focused on conflict situations exemplified by the notorious Prisoner's dilemma game.[9] Lewis was innovative in applying game theoretic models to situations in which a congruence of individuals' interests predominates over their differences. His interest in part derived from the hope of suggesting a cogent manner in which a social institution can cohere without the glue of agreement, or the threat of sanctions, otherwise required to constrain narrow self-interest.

Lewis effectively exploited John Nash Jr's game theoretic equilibrium concept. Nash argues that a stable system is comprised of a population of strategic actors, none of whom individually could enhance personal gain by choosing a different course of action, *given the acts that every other member of the population chose*. Nash's equilibrium concept is referred to as mutual-best-reply indicating that social stability is achieved at the point at which all actions offset each other in such a way that no single individual stands to gain more by acting differently. Even though Nash originally mentioned the possibility of interpreting the relevance of his equilibrium theory in a dynamic context, Lewis is credited to have been the first to use game theory to understand the outcome of games played repeatedly within a population.[10] In the 1950s and 1960s, Nash's equilibrium solution had seemed superfluous to game theorists because its formal proof provides no explanation of how actors could converge on a single mutual-best-

[2] For a recent overview of the game theoretic view of coordination and cooperation, see Peter Vanderschraaf, 'Game Theory, Evolution, and Justice,' *Philosophy and Public Affairs*, 28:4 (Autumn 1999), 325–58: 'cultural evolution produces an equilibrium…[based on a system of reciprocal expectations] that determines the community norm,' 340; for the development of the alternative position, see Margaret Gilbert, *A Theory of Political Obligation*, Oxford University Press (Oxford, 2006).

[3] See, e.g., William H. Riker and Peter C. Ordeshook, *An Introduction to Positive Political Theory*, Prentice Hall (Englewood Cliffs, NJ, 1973); Shawn Hargreaves Heap and Yanis Varoufakis, *Game Theory: A Critical Text*, 2nd edn, Routledge (London, 2004); Donald C. Hubin, 'What's Special about Humeanism,' *Noûs*, 33:1 (March 1999), 30–45, and 'The Groundless Normativity of Instrumental Rationality, *The Journal of Philosophy*, 98:9 (September 2001), 445–68.

[4] Inverse reasoning structures Rawls' fair play that was rendered senseless by mainstream rational choice theory, see his 'Justice is Political, Not Metaphysical,' *Philosophy and Public Affairs*, 14:3 (Summer 1985); for discussion see S. M. Amadae, *Rationalizing Capitalist Democracy*, Chicago University Press (Chicago, 2003), 270–3.

[5] Martin Hollis' *Trust within Reason*, Cambridge University Press (Cambridge, 1998) explores this theme from within the action type condoned by rational choice theory.

[6] John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior*, 3rd edn, Princeton University Press (Princeton, 1972), 33.

[7] Robert J. Leonard, 'Creating a Context for Game Theory,' in *Toward a History of Game Theory*, ed. E. Roy Weintraub, Duke University Press (Durham, 1992), 29–76.

[8] See Schelling's 'Reciprocal Fear of Surprise Attack,' 'Surprise Attack and Disarmament,' and 'Nuclear Weapons and Limited War,' in *The Strategy of Conflict*, Oxford University Press (Oxford, 1960), 205–66; on the general equivalence of rational deterrence theory and rational decision theory see Keith Krause, 'Rationality in Deterrence in Theory and Practice,' in *Contemporary Security and Strategy*, ed. Craig A. Snyder, Routledge (New York), 120–49.

[9] In the Prisoner's dilemma game, each prefers unilateral success gained by defection over joint cooperation, and each prefers joint cooperation over joint defection. Because each player is better off not cooperating, regardless of what the other does, the rational outcome is jointly suboptimal. See Anatol Rapoport and Albert M. Chammah, *Prisoner's dilemma: A Study in Conflict and Cooperation*, University of Michigan Press (Ann Arbor, 1965); Robert Axelrod, *Conflict of Interest: A Theory of Divergent Goals with Applications to Politics*, Markham Pub. Co. (Chicago, 1970); Richard Dawkins, *The Selfish Gene*, Oxford University Press (Oxford, 1979).

[10] For a historical overview of the priority and enduring pioneering quality of Lewis' *Convention*, see Robin P. Cubitt and Robert Sugden, 'Common Knowledge, Salience and Convention: A Reconstruction of David Lewis' Game Theory,' *Economics and Philosophy*, 19 (2003), 175–210.

reply outcome, instead of remaining at cross-purposes, in the games offering more than one solution. Lewis paved the way for considering how mutual-best-reply solutions may emerge in contexts of repeated play. Some leading theorists argue that Lewis' *Convention* still has not been fully mined for all it has to offer to contemporary social scientists studying institutions, customs, and norms.[11]

Lewis defined conventions to have two necessary qualities: they are in some sense arbitrary, and there must be a viable alternative way of organizing a practice (24). The social norms governing a convention are arbitrary in that they could have been otherwise without impairing the activity. For example, a country could determine that traffic law requires driving on either the left or the right-hand side of the road. Either is an adequate solution, so neither is definitively correct nor preferred according to an objective or non-relative standard. We may consider measuring gauges, languages, or legal practices of property rights as conventions meeting the criterion that each practice could be different, yet still effective: metric versus standard; Chinese versus German; or primogeniture versus equal inheritance to all offspring. Even though these conventions are relative, or cultural constructions, still this does not detract from them their ability to create social facts. Thus, while it is not necessary that people drive on the right, it is a fact that in the United States and most of the world, people do.

Any convention could have been otherwise. Lewis works to explain how they come about and are maintained, with two opposed possible means being explicit agreement toward, and unintended convergence onto, one of at least two possible standards. Thomas Hobbes' social contract theory is often pointed to for requiring the explicit consent of those governed to abide by the laws decreed by the sovereign. However, in the case of linguistic conventions, it seems obvious that explicit agreement to the meaning of words, in some original state of nature without language, is impossible because it would require language to convey any initial agreement. Therefore, language is a paradigmatic case of a convention that could have been otherwise equally effectively, and that must have arisen through some medium not depending on explicit agreement or consent (2).

Conventionalism was central to early twentieth-century logical positivism that had classified truths in two classes: analytic, which are conventional and self-consistent; and synthetic, which are true or false according to empirical criteria. Logical statements that merit the label "true" do so in virtue of proper relationships between terms determined by their formal linguistic framework; their truth conditions are not established by checking whether a term correctly refers to a specific empirical artifact or experience thereof. Rudolf Carnap's conventionalism is perhaps the best known for suggesting the arbitrariness of any particular linguistic framework for engaging in scientific inquiry.[12] The analytic project also extended to articulating the claim that all logical and mathematical truths are analytic, and hence devoid of any appeal to a reality underneath or beyond their formalism. The idea was to build up derivations from axioms; the axioms were stated as elemental assumptions. Gottlob Frege's logicism pioneered this basic approach. Ludwig Wittgenstein's *Tractatus Logico-Philosophicus* holds that logical truths are tautologies and are empty of content. The upshot of this line of inquiry suggests that if we accept conventions across a large range of practices, then it is at least plausible that some arise without deliberate agreement. One way to understand the direction Lewis would take is to perceive of the possibility that truths could be arrived at by convention without any act of agreement.

Quine stood against this conventionalist trend in analytic philosophy. He rejected a hermetic distinction between analytic and synthetic truths, and he argued for a holistic epistemology.[13] For Quine, even mathematics and logic are ultimately vindicated by their usefulness, and not through purely abstract or a priori analysis. Lewis takes issue with two of his mentor's tenets: first, that it is impossible to identify purely analytic truths; and second, that a basic form of assent is inseparable from either being truthful or upholding the conditions of using language truthfully.[14] In keeping with the logical positivists before him, Lewis concludes that linguistic communication is conventional. He takes his own step in arguing that linguistic meaning is a product of deploying words in regular patterns reflective of individuals' instrumental pursuits. For Lewis, the key property of an analytic truth is that it pertains to all possible worlds; how an analytic truth is expressed is purely a matter of linguistic convention. "Yesterday is past" is an example of an analytic truth that is a product of the English language, which exists by convention. Furthermore, Lewis rejects the notion that assenting to truth conditions plays any functional role in establishing conventions. Instead he proposes that linguistic conventions "are regularities in behavior, sustained by an interest in coordination and an expectation that others will do their part" (208). Lewis directly argues against Quine who looks to agreement to anchor language: "I offer this rejoinder [to Quine's argument]: an agreement sufficient to create a convention need not be a transaction involving language or any other conventional activity. All it takes is an exchange of manifestations of a propensity to conform to a regularity".[15] As will be expanded on below, Lewis' lowest common denominator that gives rise to conventions is a regularity of individuals' preferences over outcomes; these preferences are an exogenously given feature of an interactive choice context. Here he dismisses the priority of consent or agreement and replaces it with the unintended outcome of patterns of decision making governed by individuals' desires. In Quine's words, "Lewis undertakes to render the notion of convention independent of any fact or fiction of convening."[16]

Lewis' theory of language, built up from the instrumental use of signaling to achieve agents' ends, would be countered by Donald Davidson who argued that ultimately linguistic practice is prior to agents' preferences. Contrary to Lewis, Davidson held that agents are not able to form preferences over outcomes without already being part of a speech community and having linguistic mastery.[17] The game theoretic analysis of human action is wholly instrumental, and is deemed to cohere with desire-belief psychology stemming back to Hobbes and Hume. Philosophy of language and ethics are perhaps the ultimate battleground over the intelligibility of defending a fully instrumental account of action: either the type of normativity demonstrated in communication derives from agents' choices motivated by their efforts to satisfy

[11] Cubitt and Sugden, 'Common Knowledge,' 2003.

[12] The choice of a scientific framework or language is underdetermined by empirical evidence, Rudolf Carnap, 'Truth and Confirmation,' trans ed. Herbert Feigl, Wilfrid Sellars, *Readings in Philosophical Analysis*, Appleton-Century-Crofts (New York, 1949), 119–27.

[13] W.V.O. Quine, 'Two Dogmas of Empiricism,' *Philosophical Review*, 60 (1951), 20–43; for an engaging discussion of Carnap and Quine's philosophical jousting, see Thomas Uebel's entry 'Vienna Circle' in *The Stanford Encyclopedia of Philosophy*, 2006 (http://plato.stanford.edu/entries/vienna-circle/), last accessed September 16, 2010.

[14] See Lewis, 204–8, and 177–83.

[15] Lewis, *Convention*, 87–8; Cubitt and Sugden, 'Common Knowledge,' 178.

[16] Quine in Lewis' foreword, *Convention*, xii. Note Quine observes, 'The problem of distinguishing between analytic and synthetic truths was apparently one motive of the study. In the end, Lewis concludes that the notion of convention is not the crux of this distinction. He does not for this reason find the analyticity notion unacceptable, however. He ends up rather where some began, resting the notion of analyticity on the notion of possible worlds,' xii.

[17] Donald Davidson, 'Convention and Communication,' in *Inquiries into Truth and Interpretation*, Oxford University Press (1984).

preexisting desires; or linguistic competence and moral conduct may be better understood as meeting shared expectations of conduct that are independent from calculations of maximizing expected utility. Thus, according to Davidson, linguistic meaning is empty if not grounded in a predisposition toward truth-telling.[18]

Here an important philosophical divide is demarcated. On the one side are those who support the explanatory power of game theory as a reductionist program resting on the fundamental building blocks of desires that agents are presumed to rationally satisfy.[19] On the other are those who argue that truthfulness, understanding, and meaning can only be conveyed via a form of normativity that is foremost non-instrumental. This distinction in action-types can be drawn by juxtaposing one worldview in which desires are the elementary mover of all individuals' choices to another worldview in which agents' choices may be influenced by the will to conform to an intersubjective norm, rule or standard, that prescribes a choice independently from considerations of how the consequence of each choice satisfies individual preferences.[20] For example, following the rules of counting, alphabetizing and measuring enact publicly recognizable correct judgments and actions wholly distinct from how a correct answer provides a reward to the rule-follower. Davidson, Jürgen, Habermas, and Wittgenstein offer theories of meaning and normativity that are not reducible to instrumental preference satisfaction.[21] As I will discuss below, depending on which side of this philosophical divide one stands determines the resources at one's disposal for understanding action and for achieving and maintaining social order, jurisprudence, and governance. Lewis and his followers, who find in convention a constructive manner of explaining social order without recourse to extra-instrumental accounts of action, propose that they work within the realm of scientific validity without postulating the existence of entities with a dubious ontological pedigree.[22] Habermas, Margaret Gilbert and theorists not content to view all purposive and meaningful action as subsumed under the single rubric of expected utility theory suggest that the parsimony of reductionism is misguided for overlooking fundamental properties of social interaction. They further worry that this line of research may have the unintended implication of privileging and encouraging action that essentially enacts the dictum to treat others as only means and not ends.[23]

There are several charitable ways to regard the attempt to analyze social norms and conventions solely in terms of strategic rationality. First, an elegant simplicity of Occam's razor variety counsels opting for the least complex explanation. Therefore, if instrumental motives predicated on realizing desires are sufficient to understand a large variety of social practices, then this feature is a strength rather than a weakness. Second, a theory that provides an account of social organization that does not require voluntaristic adherence to norms, but only rests on rational self-interest, seems to offer a more robust explanatory strategy.[24] This way we accept up front that people are not necessarily virtuous or law-abiding. Third, it is incontestable that the advent of von Neumann and Morgenstern's *Theory of Games and Economic Behavior* provided the starting point from which to flesh out a robust scientific paradigm.[25] Given the intense and voluminous efforts that have gone into this endeavor, it has become a respected exercise to investigate the extent to which game theory can satisfactorily account for the development and stable functioning of practices and institutions throughout society.

However, even accepting these positive reasons for endorsing the game theoretic approach to analyzing norms, the fact that this approach is ultimately reductionist should be remembered so that the stakes of the ongoing debate between rational choice adherents and critics is kept in mind. Proponents of a strictly instrumentalist view of human behavior find it preposterous to step outside of a desire-belief psychology, and difficult to imagine that purposive agents can be moved by any consideration besides preference. This perspective fits with an interpretation that life is sustained through physical processes that have spatio-temporal locations.[26] Lewis subscribes to Humean supervenience holding that every existing phenomenon supervenes on the distribution of properties embodied by space-time points. Preferences are reflected in brain states and play a vital role in propagating life. Along this line of thinking, only instrumental normativity is coherent.[27] It is self-validating because acts expressing instrumental consistency are rewarded by success. On the other hand, philosophers who propose categories of action that are non-instrumental but not irrational suggest that some practices are governed by regularities in behavior that result from judgment and action for which there is no tidy correlation between appropriate choice and desire gratification. Possible candidates for non-instrumental normative action can be found in Wittgenstein's account of language games in which rules guide action, and not a calculation of how the action will maximize expected utility. Furthermore, it seems a stretch to suggest that the practice of mathematics is guided by desire gratification instead of by a commitment to understanding and developing mathematical truths, unless one postulates that preference fulfillment is directly correlated to uncovering truths.[28]

Understanding Lewis' *Convention*

Given the widespread mid-century appeal of a conventional approach to philosophy, it is not surprising that David Lewis had the insight of treating social institutions from politics to linguistic mastery as conventions that are contextually relative. But he went further than working within the philosophical legacy of logical positivism. He investigated the possibility of accounting for conventions independently from agreement, consent, assent,

[18] Donald Davidson, 'The Folly of Trying to Define Truth,' *Journal of Philosophy*, 93 (1996), 263–78; 'The Structure and Content of Truth,' *Journal of Philosophy*, 87 (1990), 279–328; 'Truth and Meaning,' in *Inquiries into Truth and Interpretation*, Oxford University Press (1984); for discussion, see Joseph Heath, *Communicative Action and Rational Choice*, MIT Press (Cambridge, MA, 2001), 19–22. As Heath explains, for Davidson it is not that a desire for truth is necessary for linguistic communication, but rather than truthfulness is a necessary precondition for desire, 22. For Lewis, desire is prior to communication and truth.

[19] For this vocabulary, see Heath, *Communicative Action and Rational Choice*, 79.

[20] For a helpful discussion see Heath, *Communicative Action and Rational Choice*, 1–10.

[21] Davidson, 'Truth and Meaning'; Habermas, *Theory of Communicative Action*; Ludwig Wittgenstein, *Philosophical Investigations*, trans. ed. G.E.M. Anscombe, 2nd edn, Blackwell (Oxford, 1997).

[22] In 'An Argument for Identity Theory,' Lewis makes the case for maintaining explanations of experience within the laws of physics, *Journal of Philosophy*, 63 (1966), 17–25.

[23] This is Habermas' concern in his thesis that systems of non-discursive interaction are colonizing the lifeworld of shared meanings and practices, *Theory of Communicative Action*, 2 vols., trans. ed. Thomas McCarthy, Beacon Press (Boston, 1984–87).

[24] This reason is typically cited as buttressing the rational choice approach, see e.g., James M. Buchanan and Tullock, *Calculus of Consent*, University of Michigan Press (Ann Arbor, MI, 1962), and Buchanan's *Limits to Liberty: Between Anarchy and Leviathan*, University of Chicago Press (Chicago, 1975).

[25] Amadae, *Rationalizing Capitalist Democracy*, 1–14.

[26] Note that apparently Lewis rejected the quantum mechanical Bell inequalities that suggest action at a distance, Brian Weatherson 'David Lewis,' *Stanford Encyclopedia of Philosophy* (http://plato.stanford.edu/entries/david-lewis/), last accessed September 16, 2010.

[27] For a defense of the unproblematic quality of instrumental rationality see Donald C. Hubin, 'What's Special about Humeanism,' and 'The Groundless Normativity of Instrumental Rationality.'

[28] For discussion of the source of mathematical normativity see Crispin Wright's *Wittgenstein and the Foundations of Mathematics*, Duckworth (London, 1980); see also Stewart Shapiro, *Thinking about Mathematics*, Oxford University Press (Oxford, 2000).

commitment, or promising. Whereas a covenant confers binding force specifically because agents pledge allegiance to its stipulations, conventions are products of an alignment in individuals' interests that converge into a mutually beneficial outcome from common knowledge of mutual expectations. Lewis relied on the non-technical application of game theory to devise a comprehensive theory of convention. In order to do this, he worked to associate an intuitive understanding of convention with the game theoretic definition of coordination problems. Superseding the 1950s and 1960s trend to study conflict, Lewis made his name by redirecting attention to games in which individuals' interests align such that everyone favors achieving the same, instead of different outcomes. Lewis' effort to save a conventional account of analyticity, and his move to explain linguistic communication without agreement, was made possible by drawing on instrumental normativity as expressed in the purely analytic rational decision theory. Lewis confronted the dilemma of how agents can coordinate their actions in a linguistic or civil community without depending on any source of normativity besides maximizing rational self-interest.

In the late-1960s, when Lewis applied game theory to his theoretical investigation of the philosophy of language, this approach was wholly undeveloped. Thus when Lewis cites Schelling's *Strategy of Conflict* for providing the point of departure for his own analysis, he makes clear that the original distinction between games of conflict and coordination existed as a matter of family resemblances and not formal definition. Conflict situations are those in which agents seek different outcomes in competition with each other. The term "zero-sum" derives from game theory to specify an interaction in which there is an objective or intersubjective means of keeping score, and one person's gain is necessarily another's loss. In circumstances of coordination, agents would rather arrive at the same outcome. Schelling's initial characterization is schematic:

> If *chess* is the standard example of a zero-sum game, *charades* may typify the game of pure coordination; if *pursuit* epitomizes the zero-sum game, *rendezvous* may do the same for the coordination game.[29]

In pure conflict games, agents' preferences over outcomes are inversely correlated; in pure coordination games, their preferences are positively correlated. Many games have both elements, thereby representing mixed game forms that may be classified as mainly conflictual or mainly coordinating.[30]

From the game theory perspective, the decisive feature of all games ranging from pure conflict to pure coordination is that of strategic action: each agent's estimation of a best course of action is dependent on knowing or expecting what others will do. Schelling explains the essence of strategic rationality:

> It is a behavior situation in which each player's best choice of action depends on the action he expects the other to take, which he knows depends, in turn, on the other's expectations of his own. This interdependence of expectations is precisely what distinguishes a game of strategy from a game of chance or a game of skill.[31]

There is a specific assumption in Schelling's exposition of coordination that epitomizes the game theoretic approach to communication and prefigures Lewis' theory of language. The fact that actors are strategic, treating others as elements of an environment for maximizing personal expected utility, is primary. In cooperation or conflict, agents' strategic actions depend on their rational expectations of what course of action others will take. These expectations are more fundamental than communication in determining which act to choose. Whether an interaction circumstance is more conflictual or coordinating is a function of the joint profile of all participants' preference orderings over potential outcomes. As Schelling points out in the context of bargaining, it may not even be helpful to an individual to make his own preferences transparent to his trading partner.[32] Schelling is explicit,

> [T]alk is not a substitute for [game] moves. Moves can in some way alter the game, by incurring manifest costs, risks, or a reduced range of subsequent choice; they have an information content, or evidence content, of a different character from speech. Talk can be cheap when moves are not (except for the "talk" that takes the form of enforcible threats, promises, commitments, and so forth, and that is to be analyzed under the heading of moves rather than communication anyway).[33]

The point is that in social situations of all kinds, individuals are strategic actors seeking to realize their preferences over outcomes as a function of their expectation of the actions everyone else will take. Of course, others' choices are similarly dependent. Lewis follows Schelling in specific, and game theory more generally, in holding communication to be one agent's act to send a signal to someone else, calculated to maximize one's expected utility by producing a response in the other. Lewis also concurs with Schelling in drawing attention to the crucial role of mutual expectations which ultimately become formally defined as common knowledge.[34] In game theory, communicating only serves the strategic function of furthering one's interests, and can have no auxiliary or additional power. If meaning is conveyed, this is secondary to the elementary fact that individuals use signals to achieve their ends.

Lewis draws on Schelling's *Strategy of Conflict* to respond to Morton White and Quine's theories of language. Quine observed in his introduction to *Convention* that the implications of Lewis' analysis span from language to the social contract. The central puzzle is how linguistic meaning or laws of governance could have been established given that it is generally acknowledged that neither were nor could have been the product of assent or agreement. Lewis claims that his theory is consistent with David Hume's view on the "origin of justice and property" that develop as regularities in behavior produced by underlying interests people have.[35]

Lewis introduces eleven contexts of coordination that are socially diverse, from Jean Jacques Rousseau's heuristic stag hunt, to interrupted telephone conversations, and the selection of a common exchange currency. Obviously, how an individual selects to act is dependent on how he believes others will act. Lewis is quick to introduce John Nash's mathematically formulated equilibrium concept to identify potential solutions to interaction problems in which agents' preferences over outcomes are positively correlated.[36] A game theoretic solution to these problems stipulates that a stable, or equilibrium, outcome is

[29] Schelling, *Strategy of Conflict*, 85.

[30] For Schelling's discursive typology, see *Strategy of Conflict*, 89; for Lewis' discussion of this typology, see *Convention*, 13–5.

[31] *Strategy of Conflict*, 86.

[32] *Strategy of Conflict*, 116.

[33] *Strategy of Conflict*, 116.

[34] Cubitt and Sugden, 'Common Knowledge,' and Robert J. Aumann, 'Agreeing to Disagree,' *Annals of Statistics*, 4, 1236–9.

[35] Lewis, *Convention*, 3–4.

[36] It is interesting that Lewis does not give Nash credit for the equilibrium concept he uses, *Convention*, 8.

achieved if in that state it behooves no individual to choose an alternative action given what everyone else chose to do. This is referred to as a mutual-best-reply set of choices that are self-reinforcing insofar as if the scenario were to be repeated, chances are nobody would diverge from her prior choice. This follows from the fact that no one could deviate and achieve a higher payoff, given what everyone else did. However, crucial for game theory and for Lewis' analysis, there may be multiple such Nash equilibrium outcomes. Moreover, some equilibrium outcomes may be preferable to others from different agents' viewpoints. Thus, in the case of multiple possible equilibrium outcomes that Lewis deems characteristic of the coordination problems underlying conventions, it is unclear how agents converge on a single mutually acceptable outcome. For example, in a rendezvous predicament in which agents are not sure which of two locations to meet, but both are eager to see each other, each agent must decide where to go given his anticipation of the other's decision, vice versa, and further in consideration of the other's knowledge of oneself. Therefore a simple meeting problem can be interpreted as an infinite regression of mutual expectations.[37]

There are several dimensions to grasping Lewis' concept of convention which he defines as "situations of interdependent decision by two or more agents in which coincidence of interest predominates and in which there are two or more proper coordination equilibrium" (24). In short, agents' preferences must be aligned; they must share mutual expectations; and there must be more than one viable coordination equilibrium. On the one hand, there is an intuitive sense, and on the other there is his attempt to provide a relatively formal definition drawing on game theory. However, Lewis does not provide an axiomatized system to pinpoint his concept of convention. Therefore, even though convention bears a somewhat close relationship to Nash's axiomatically defined equilibrium, it remains challenging to grasp Lewis' meaning with perspicacity.[38] An admirer of Lewis defends the philosopher's abundant sprinkling of "almost" throughout his definition of convention. Apparently Lewis believed that because this term characterizes a "vague" "folk concept" its analysis should similarly be vague and imprecise.[39] Lewis comes to finally provide the following definition:

> A regularity *R* in the behavior of the members of a population *P* when they are agents in a recurrent situation *S* is a *convention* if and only if it is true that, and it is common knowledge in *P* that, in almost any instance of *S* among members of *P*,
>
> (1) almost everyone conforms to *R*;
> (2) almost everyone expects almost everyone else to conform to *R*;
> (3) almost everyone has approximately the same preferences regarding all possible combinations of actions;
> (4) almost everyone prefers that any one more conform to *R*, on condition that almost everyone conform to *R*;
> (5) almost everyone would prefer that any one more conform to *R′*, on condition that almost everyone conform to *R′*.
>
> where *R′* is some possible regularity in the behavior of members of *P* in *S*, such that almost no one in almost any instance of *S* among members of *P* could conform both to *R′* and to *R* (78).

The gist of this definition is that practices ruled by convention depend on uniformity of agents' preference that everyone conform, and common knowledge that everyone expects everyone to conform. Assuming that agents' preferences over end states are elemental, and that agents seek to maximize personal preference satisfaction in situations with other like agents who are transparent insofar as individuals share knowledge of each other's preferences, the Nash equilibrium identifies solution points. These are outcomes that are a product of individuals' maximizing behavior, and are characterized by stability if attained because no single individual can augment personal gain by deviating from the course of action selected by all the other agents.

The Nash equilibrium has the weakness of not providing a rationale by which it would be arrived at in the first place. If there are multiple equilibria, how do agents gravitate to one, rather than end up in an uncoordinated state as individuals each tries to realize a different outcome? However, as game theorists analyzed recurring situations that provide agents with the opportunity to revise individual action choices in response to what other agents did, they proposed that over some extended period of repeated play, presuming that agents' preferences remain constant, that there will be convergence upon a Nash equilibrium. Still it is not clear in games having more than one equilibrium, how this process moves toward one stable outcome instead of another. Lewis' conventions are Nash equilibrium, but with the addition of two more rarified criteria: at least two possible equilibrium exist, and a "coincidence of interest predominates" (24).

Intuitively, Lewis' coordination equilibrium is a more restrictive definition than that of Nash's mutual-best-reply because he requires not only that no individual can achieve a better outcome by personally deviating from what everyone else chose to do, but moreover that no individual could be better off if any other individual had adopted a different course of action.[40] Thus, in mutually amicable coordination situations, people prefer the same outcome as each individual benefits from his own participation, and from the participation of (almost) everyone else.[41] In problems of conflict, such as the notorious Prisoner's dilemma, a Nash equilibrium may exist. However, it is not a convention because, although no individual can improve on his condition through another course of action, each individual could be better off had his opponent chosen to cooperate instead of to defect. For example, in warfare one side may have been as successful as possible given the opponent's bombing strategy, but would have been more successful if the opposition had surrendered.

Accepting, then, that individuals prefer to converge on a single outcome in coordination problems, it may seem that this area of philosophy is trivial because agents face no real obstacles in this class of predicaments. If everyone is better off by arriving at a mutually preferred outcome, and no one can do better by unilaterally taking a different action or by goading any other individual into taking a different action, it may seem puzzling that anything remains to be investigated. To some extent Lewis had the same worry as he worked to eliminate from discussion trivial coordination problems: he is specifically interested in situations in which there are multiple possible equilibria, and it is not clear how agents will alight on the same outcome rather than inadvertently acting at cross-purposes.[42] Lewis must also dismiss agreement as superfluous, even though it seems to offer an intuitively obvious means for agents with aligned interests to coordinate their actions without ambiguity.

Lewis' exposition of convention moves between discursive discussion and figurative display of game theoretic matrices. He

[37] Lewis, *Convention*, 27–32.

[38] See Margaret Gilbert, 'Game Theory and Convention,' *Synthese*, 46:1 (January 1981), 41–93, especially 47.

[39] Brian Weatherson, 'David Lewis,' *Stanford Encyclopedia of Philosophy*, 5 (http://plato.stanford.edu/entries/david-lewis/), last accessed September 16, 2010.

[40] For a helpful discussion, see Michael Rescorla, 'Convention,' *Stanford Encyclopedia of Philosophy*, September 6, 2007 (http://plato.stanford.edu/entries/convention/), last accessed September 16, 2010.

[41] Lewis maintains the caveat that a convention may exist even if almost everyone participates.

[42] For Lewis' elimination of trivial coordination problems, see *Convention*, 69–76; for discussion see Gilbert, 'Game Theory and Convention,' 54–75.

follows his eleven examples to illustrate how each can be resolved by a convention that reflects a general alignment of interests, and yet is not trivially reducible to only one outcome.[43] Drivers could drive in the right or left lane. In the case of an interrupted phone call, either the call initiator or receiver could call back. Two people rowing a boat could have any of a number of rhythmic stroking patterns. People could meet regularly every week at one of several meeting places. Members of an oligopoly could set prices in any number of ways to their advantage. In Rousseau's proverbial stag hunt, agents could work together to hunt stag, or they could go it alone to hunt hare. People could adopt any of a large number of media of exchange, from gold to symbolic currency. People can obviously effectively speak different languages and dialects. Lewis argues that the regular pattern of individuals' preferences structuring each of these interaction contexts may result in an ongoing stable convergence onto one particular outcome with the additional factor of agents' common knowledge of a preexisting convention. The rowers continue to stroke in a certain rhythm that is already established and is to everyone's advantage. Lewis consistently rejects that agreement on an outcome is necessary or even helpful for motivating a joint plan of action as the explanation for how conventions are maintained.

Lewis' denial of agreement as the force sustaining conventions provides two different reasons that social scientists may have been attracted to his method of analysis. As was discussed at the outset, to some extent Lewis set out to provide an explanation for how language may exist that does not have recourse to the impossibility of resting the explanation on prior agreement. Given that his final two chapters address linguistic communication, one could read his earlier chapters, which pin down the conceptual apparatus of conventions, as laying the groundwork for providing a coherent account of language that recognizes analytic truths yet does not presuppose agreement. However, Lewis' disinterest in the potential role that agreement could play in giving rise to and maintaining conventions is also relevant to social scientists who consider instrumental action to be more basic than communication. In game theory, even though it may be possible to identify a solution to a game based on agreements that are backed by external force, non-binding agreements are typically viewed as little more than 'cheap talk.' Agents will say whatever they deem will further their purposes. Thus, social scientists find in Lewis' convention a way to explain activities that may seem to be the product of a joint or community goal, but rather can be explained by understanding the convergence of individuals' preferences in such a way that once a convention is commonly known to be established, it is in no single individual's or small group of individuals' self-interest to take a deviant course of action.

By Lewis' analysis, agreement is not necessary to initiate a convention, nor to perpetuate it. In fact, Lewis goes further in using his definition of convention to rule out the usefulness of agreement for generating any binding cohesion that produces the regularities we observe in social institutions. He observes, "a convention begun by agreement may not become a convention, on my definition, until the direct influence of the agreement has had time to fade" (84). Specifically, Lewis seeks to eliminate the role agreement may be thought to play in creating the mutual expectations underlying normative practices. He shifts the operative element from the role of agreement in shaping an individual's commitment to conform to a norm by artificially modifying her preferences despite what others do, to the straightforward motivating role the individual's preferences have on determining her actions. He explains,

> Suppose we all swore a solemn and public oath to conform to *R* [some suitable regularity] come what may. Then for a while we might all prefer *un*conditionally to conform to *R*, each determined that even were the others to break their oaths and conform to some alternative regularity *R′*, still he would rather keep his oath (84).

Basing a social practice on an oath, according to Lewis, runs counter to the central idea of convention because, in order to move an agent to act, oath swearing renders the emergent practice non-arbitrary in the sense that only this practice is deemed legitimate. Moreover, in Lewis' analysis, taking a promise gives one the preference to uphold it *regardless* of what other agents do in contrast to conventional actions in which it behooves one to comply because one anticipates that all or most of the others will as well.

To recap Lewis' account of convention, he supplies the following 5 propositions to indicate the basis of a conventional practice if all are true:

(1) Most other members of *P* involved with me in situation *S* will conform to *R*.
(2) I prefer that, if most other members of *P* involved with me in *S* will conform, then I conform also.
(3) Most other members of *P* involved with me in *S* expect, with reason, that I will conform.
(4) Most other members of *P* involved with me in *S* prefer that, if most of them conform, I conform also.
(5) I have reason to believe that (1)–(4) hold. (97)

Again, conventions are maintained by consistent preferences throughout a population that one generally beneficial outcome be brought about, and everyone's mutual expectation that everyone is motivated to help bring it about. The convention exists because everyone profits from its enactment, and because everyone bases his choice on the expectation that others will choose similarly. Having claimed that these propositions characterize a conventional practice, Lewis goes on to deduce two further claims. The first of these is crucial for realizing that the normative force of a convention derives from individuals' preferences over outcomes.

(6) I have reason to believe that my conforming would answer to my own preferences (98).

Lewis emphasizes, "we do presume, other things being equal, that one ought to do what answers to his own preferences" (98). The pivotal point in Lewis' argument is that in institutions built on conventions, individuals ought to conform not due to some ancillary action of promising, and not due to a moral 'ought.' Instead individuals go along because complying with the regularity of the practice, once embodied in individuals' mutual expectations, is in one's best interest. The normativity of a conventional activity is strictly the 'ought' of the instrumental imperative to satisfy one's desires.

Lewis adds proposition number seven:

(7) I have reason to believe that my conforming would answer to the preferences of most other members of *P* involved with me in *S*; and that they have reason to expect me to conform (98).

Thus, it may seem that Lewis introduces either a moral concern for treating others as ends and nor just means, or altruistic preferences. However, Lewis only suggests that if an action maximizes one's own expected utility, then the fact that it also

[43] Lewis, *Convention*, 42–51.

promotes others' maximization of expected utility can only add to the instrumental validity of this choice of action.[44]

Lewis' conventionist and instrumentalist account of the normativity underlying the social contract and linguistic truthfulness

Lewis concludes his text with chapters on communication and language, which occupied his interest more than the implications of convention for political philosophy of the social contract, and takes a firm stand on the ultimate instrumentalism of linguistic performance and meaning. As has become evident in Habermas' *Theory of Communicative Action*, linguistic normativity may be directly related to understanding sources of legal and democratic legitimacy.[45] For Habermas, linguistic dialogue grounds social normativity.[46] Lewis' argument tack is to first show that social practices including governance depend on interactive regularities, and then to propose that language is conventional and instrumental. In chapter three, Lewis contrasts convention with agreement, social contracts, norms, and rules. Regardless of the merits of Lewis' argument regarding the foundations of communication, political theorists have found his exposition on conventions versus social contracts useful for analyzing political interactions in terms of interests as opposed to prospectively extra-instrumental practices such as promising, tacit consent, or fair play.[47] Lewis' *Convention* initiated this trend of research that continues to be popular well into the twenty-first century. However, I raise the question of whether Lewis is ultimately successful in reducing the normativity of linguistic truthfulness and of abiding by the social contract to strict instrumentalism.

Lewis asks directly, ''Is my concept of convention nothing but our familiar concept of social contract, as inherited from Hobbes, Lock, and Rousseau, demythologized, and applied to matters other than political allegiance and social solidarity?'' (88). He is quick to answer, ''It is not. The concept of social contract, as I understand it, is different in principle from that of convention'' (89). Lewis differentiates between conventions and contracts by stipulating distinct definitions that could, in particular games, overlap, but need not. A convention is characterized by a confluence of interest such that I want to go along with whatever everyone else does. A social contract, by contrast, has two polar extreme outcomes such that one is suboptimal, as is Hobbes' proverbial state of nature, and one is roundly preferred to the suboptimal outcome. Furthermore, social contract games tend to have the form that some or many agents prefer to be the lone defector or free rider. Thus, according to Lewis, in social contracts it is typical that I prefer others to abide by the contract while I do not; conventions have the form that I prefer both I and others abide. If a social contract is a convention, then there must be at least two possible mutually likeable world states, and everyone prefers that everyone conform, including himself.

We gain three key insights from Lewis' contrast of contracts versus conventions. First, in some cases a situation may arise that satisfies both definitions. Let us suppose that in some population, individuals all prefer to voluntarily pay taxes and engage in voluntary public service. In this case, both the criteria of a convention, that everyone prefer himself and everyone else to comply, and of a contract, wherein the bad equilibrium is total non-compliance, and the good equilibrium is voluntary acceptance of civic responsibility, hold. The key point is that agents all prefer voluntary cooperation without the intercession of a pledge to this effect that would transform selfish preferences into civic-mindedness. Lewis rejects the practice of pledging as a plausible basis for convention because he takes it to require that individuals adopt unconditional allegiance to a rule of conduct, regardless of whether others comply. If these conditions are met, then for Lewis convention and social contract can overlap in some circumstances:

> If we return to our ordinary, wider concept of preference [that encompasses ethical considerations], it remains true that many social contracts will be sustained by the moral obligations of tacit consent or fair play, as recognized by the agents involved. But these accepted obligations will be counted as a component of preferences, not as an independent choice-determining force (94).

Presumably Lewis has two points in mind for his conventionalist interpretation of the social contract to hold. He disavows the relevance or necessity of duty, or a form of action external to preference satisfaction. As well, he suggests that individuals' voluntary adoption of moral obligation is contingent on others likewise adopting such obligation. Lewis accepts that tacit consent, fair play, or obligation, insofar as any move individuals, is reflected in individuals' preference to abide by the social contract.

Second, Lewis accepts that a social contract must rest on obligation given the understanding that agents who accept political responsibility have integrated moral considerations into their preferences. ''So our social contract is a convention after all. But it is a convention because of the modification of our preferences by obligations, and these obligations exist because it is a social contract'' (94).[48] Third, ultimately, though, Lewis rejects many social contracts as satisfactory cases of a convention because he finds they do not comport with the stipulation that conventions have at least two roughly equally good alternatives. Given the two polar extremes of civil society and state of nature, Lewis finds that the social contract may not qualify as a convention (95). Lewis argues that Hobbes' social contract, which only has two stark choices, is not a convention because there are not at least two relatively equally compelling outcomes to select from. Yet he claims that Rousseau's social contract based on the stag hunt does qualify because hunting stag or hare are both plausibly satisfactory (95).[49] Most likely this final consideration will not be found relevant because most societies could be governed by a variety of constitutions and rules.

Before examining Lewis' success in reducing normativity to instrumental action, it is worth pausing to observe the following. Lewis implies that any social contract, at least insofar as it is conventional and not governed by de facto force alone, rests on tacit consent or fair play. It is important to note that Lewis insists that the definition he ''gave of convention did not contain normative terms: 'ought,' 'should,' 'good,' and others.'' Therefore

[44] Lewis seems to suggest acting in accordance to economists' Pareto condition stipulating that if an action makes at least one person better off and no one worse off, it is desirable and respects a principle of minimal benevolence; it provides the slimmest endorsable moral ought that comports with a philosophical commitment to respecting individuals as expected utility maximizers.

[45] Habermas, *Theory of Communicative Action*, and *Between Facts and Norms*, MIT Press (Cambridge, 1996), and Heath, *Communicative Action*.

[46] See Habermas, *Theory of Communicative Action*, vol. 1 and 2.

[47] For the different structure of each of these, see A.J. Simmons, *Moral Principles and Political Obligation*, Princeton University Press (Princeton, 1979).

[48] Lewis seems to imply that this permissible transformation of preferences underlying a contract is not of the active sort required by intentionally agreeing, consenting, or promising.

[49] On this point, Margaret Gilbert points to a contradiction in Lewis' argument because by his own definition the stag hunt cannot qualify as a social contract. This follows because individuals who choose the certainty of hare over the possibility of stag are moved by the consideration of a maximin security threshold that is wholly independent from making a decision based on what one expects others to do. For Lewis, that everyone chooses hare must follow from mutual expectations, and not from the extraneous consideration that lone hare hunting is a safe choice *regardless of what others choose to do*; 'Game Theory and Convention,' 51–4.

it is the case that "'convention' itself. . .[on Lewis' analysis], is not a normative term" (97). In this set of pages we glean that Lewis accepts that only preference and desire motivate agency, and acknowledges that only instrumental rationality can generate valid 'ought' claims. He holds the view that "we do presume, other things being equal, that one ought to do what answers to his own preferences" (98). He observes that

> If we think of someone's preferences as the resultant of *all* the more or less enduring forces that go into determining his choices, then action that regularly goes against preference is barely possible (93).

In short, in keeping with a fully instrumentalist understanding of agency of which expected utility theory is one species, any factor registering in a agent's choice of action must inform her preferences over outcomes.

For Lewis' argument to hold, the moral obligation to act in accordance with the social contract rather than lone defection must be instrumentally derived from preference satisfaction. He seeks to rule out that norms and action conforming to them could exist as behavior governed by extra-rational moral obligation. He notes that it is true that "Sometimes, however, we think of preference more narrowly as the resultant of choice-determining forces *other than* a sense of duty. Our accepted moral obligations can and do regularly override our preferences in this narrow sense" (93). However, the key point is that "these accepted obligations will be counted as a component of preferences, not as an independent choice-determining force" (94). It is crucial that Lewis finds it necessary to distinguish between narrowly self-interested preferences, and more fully expressive preferences that might include other-regarding or moral considerations. Note, too, that on Lewis' account, recourse to sanctions to entice law-abiding behavior fails to achieve a convention because, as with pledges, agents may well comply regardless of what others do.

Lewis next develops his conventionalist explanation of language. Although I cannot do justice to it here, he challenges H.P. Grice's and John Searle's views that intent to communicate is a necessary attribute of linguistic performance (152–9). Lewis states, "I have been arguing that once we capture the conventional aspect [of language], we are done. We have captured the intentional aspect as well." (159) Any intention is already captured in instrumental preference satisfaction; no recourse to an additional concept of action type is required.

Towards the end of *Convention*, Lewis arrives at the often held position that a relation holds between linguistic communication and civil society. The normativity underlying social cohesion is viewed on par with the normativity of linguistic use. For Lewis, the same conventional account resting on instrumental normativity is valid for both. The only relevant and operational 'ought' is that of expected utility maximization.

One indispensable feature of effective linguistic communication is that people use signals that have a conventional meaning in a truthful way. Here the parallel to Lewis' earlier conventionalist account of the social contract is made explicit. Lewis observes, "A convention of truthfulness in L [a possible language] is a social contract as well as a convention" (182). Lewis confronts the issue that even though people mainly have a direct interest in being truthful to achieve their ends, nonetheless one could imagine that on occasion one may prefer to be a lone defector from a convention of truthfulness in order to realize one's interests.

Lewis' reasoning is the same here as in the earlier chapter specifically addressing social contracts of government. Just as people prefer to live in civil society over a state of nature, they prefer to be part of a linguistic community rather than living in Babel, the linguistic equivalent to every man standing for himself. Even though Lewis does not belabor his case that language is a social contract, in making his argument, Lewis renders explicit the challenge of locating all normativity in instrumental preference satisfaction. According to Lewis, language is a social contract because:

> Not only does each prefer truthfulness in L by all to truthfulness in L by all but himself. Still more does each prefer uniform truthfulness in L to Babel, the state of nature. *So each ought to recognize an obligation of fair play to reciprocate the benefits he has derived from others' truthfulness in L, by being truthful in L himself* (182, emphasis added).

The definition of a social contract is that individuals prefer uniform conformity to a state of nature, but there is the possibility that making an exception for oneself is preferred to uniform conformity. Lewis eliminates this possibility by making the case that a predilection for lone defection over uniform conformity is contrary to establishing a viable convention. Hence individuals must, he argues, adopt the predisposition to abide by linguistic conventions, even when on occasion this may counter self-interest.

Lewis elaborates:

> This much is true: one who is truthful in L against his own preferences cannot then be acting in conformity to a convention. But such cases are exceptional. In the world as we know it – and as it must be, if use of language is to persist among sinful men – almost everyone almost always has reason to get others to share his beliefs, and therefore has reason to conform to conventions of truthfulness. Thus, in the normal case, one can both be fulfilling a moral obligation and be acting according to one's preferences (182).

Here his reasoning mirrors his earlier discussion that a social contract does not require promising or agreement. Previously he insists that action is consistent with preferences, and that it is possible, even reasonable, that people adopt the moral predilection to prefer uniform conformity to personal lone defection. There is a slight slippage here because Lewis implies that a generalized commitment to the status quo over a state of nature necessarily entails adopting the preference that the status quo is superior to lone defection. But, of course, this entailment cannot be deduced from strict instrumentalist considerations. Just because one overall likes to live in a civil society with language does not imply that individuals either prefer to forgo, or actually will forgo, the occasional opportunity to defect by cheating or lying. At least, this is the predominant worry throughout the rational choice canon.[50] Although this gap in argumentation is not overt in Lewis' earlier discussion of the social contract, it becomes clear in his treatment of truthfulness in language.

We could voluntarily take on a moral commitment to truthfulness and for the most part, being truthful will directly serve our interests. However, Lewis admits that there are "exceptional cases" in which it may better serve our preferences to defect from the convention, hence calling into question the viability of the convention which is defined by unproblematically fulfilling individuals' preferences. Lewis simplifies the decision problem by postulating that in a convention individuals prefer joint coordination to unilateral defection and that in a social contract individuals prefer joint cooperation to joint defection, then posing the stark question of whether individuals prefer joint coordination, joint defection, or unilateral defection. He presents

[50] See, e.g., Russell Hardin, *Collective Action*, Resources for the Future (Washington DC, 1982).

an all or nothing case for always coordinating, or always defecting, which would of course place lone defection on par with a state of nature if sufficient people select that option. The choice is more accurately construed to be between uniformly complying, or unilaterally defecting when occasionally in one's self-interest. The role of fair play and moral obligation is specifically to forestall defection in the exceptional cases. If defecting in the case of lying or cheating were truly exceptional in a given community, then doubtlessly linguistic communication and governance would not be threatened. However, if individuals' practice were to lie and cheat each time an opportunity arose that directly fulfilled their preferences, then it likely follows that truth-telling and law following would be challenged and would collapse as conventions.

Lewis accepts the necessity of moral obligation playing a role both in the social contract and linguistic truth-telling. This obligation is on par with John Locke's tacit consent or John Rawls' fair play. Lewis introduces the decidedly normative vocabulary of 'ought': "each ought to recognize an obligation of fair play to reciprocate the benefits he has derived from others' truthfulness in L himself" (182). The decisive point is that in the exceptional cases, "*this obligation will be his only reason to be truthful in L*" (182). Both the governmental social contract and linguistic communication require the moral obligation to abide by the convention in exceptional cases (93, 182). Individuals' voluntary adoption of this preference transformation to accept moral responsibility is insufficient to guarantee that uniform compliance is preferred to lone defection once all the exceptional cases are also considered. Therefore Lewis seeks to provide an instrumental rationale to justify this preference transformation. He simply deduces that abstaining from exceptional lone defection on the occasions that it suits one's interests is necessary given the overall preference for civil society and communicative sociability. Thus, Lewis concludes, "his obligation arises because he prefers the status quo to the state of nature...the conditions of such obligations are present for everyone" (94). Hence in any social contract, people must act on moral obligation contrary to their narrow preferences at such times that only this felt obligation provides a rationale for acting.

Lewis comes close to accepting that the basis of civil society and the convention of linguistic truthfulness, although generally in individuals' self-interest, rests on the acceptance of moral obligation to abide by generally accepted standards of conduct independent from individuals' preferences. Of course, moral obligation, once accepted, may be regarded as "a component of preferences, not as an independent choice determining force" (94). However, the salient point is that, although necessary for the social contract as a convention to exist, this moral obligation does not flow from preferences. Lewis observes, "Hence, for any social contract, the condition of such obligations are present for everyone. If everyone will recognize such obligation, everyone will honor the social contract *whether or not he prefers to*" (94, emphasis added).

The question is, has Lewis violated his own attempt to purely derive the conventions requisite for social order, the social contract and linguistic truthfulness, from instrumental preference satisfaction? Has Lewis managed to walk the fine line between accepting that society relies on normativity, but rendering all normativity an expression of instrumental fulfillment? At the point that Lewis invokes wider versus narrower preferences that follow from the voluntary adoption of moral obligation underlying tacit consent or fair play, the issue may seem purely semantic. Moral obligation can counter narrow preferences, but is still an expression of preferences (93). However, Lewis attempts to claim more than that some individuals may feel moral obligation to abide by the social contract. He claims that action in accordance with the social contract and linguistic truthfulness is required for these practices to exist. Thus he attempts to derive voluntary obligation from pure instrumental logic: everyone's "obligation arises because he prefers the status quo to the state of nature" (94). One ought to prefer not to be the lone defector because, according to Lewis, this is a condition of exiting the state of nature.

Lewis admits that his reasoning has the same structure as Rawls' argument for fair play (93–4). Rawls essentially argues that if a person accepts that society's basic rules reflect his interests overall, then this individual will have a reason to comply with the law even if this compliance will not directly fulfill his interests on a case-by-case basis. Fair play, then, relies on a type of commitment or acceptance of responsibility that voluntarily yields momentary self-gain to the consideration that an individual agrees that this set of rules is sound. In Rawls' words, "if the participants in a practice accept its rules as fair, and so have no complaint to lodge against it, there arises a prima facie duty...of the parties to each other to act in accordance with the practice when it falls upon them to comply."[51] Doubt remains that Lewis is successful in fully instrumentalizing the normativity that even he admits is required to maintain government and linguistic truthfulness. The rational choice community roundly decided that Rawls' fair play is not consistent with the minimalist concept of expected utility maximization.[52] Rawls eventually concurred with their estimation of the implications of his philosophical position. Rawls agrees with Lewis that individuals have a moral obligation to abide by rules they deem appropriate rather than act as lone defectors. However, he ultimately finds that this motivation is derived from tacit agreement to standards of conduct, and from internalizing them as guides to action that override the instrumental drive to satisfy preferences. He explains, "[a]s with any moral duty, that of fair play implies a constraint on self-interest in particular cases, on occasion it enjoins conduct which a rational egoist strictly defined would not decide upon."[53] Rawls argues that in the exceptional cases, it is not possible to deduce the necessity to abide by the communal standards of action from the momentary maximization of expected utility.

Lewis has the following avenues open to him to prevail in his argument that the conventions maintaining society can function purely on instrumentalism. He could rest his case on the sufficiency of people's self-interest to usually comply. In fact, he anticipates this solution by his careful inclusion of the caveat 'almost' throughout his definition of convention; almost everyone finds it in his interest to coordinate. Yet, I think that he is right to suspect that the social contract and linguistic truthfulness require more than a predisposition to cheat or lie on a case-by-case basis decided by instantaneous calculation. Social cohesion functioning as a convention is compromised by a widespread predilection for opportunistic unilateral defection determined by momentary cost-benefit analysis. This worry is of course the standard worry for political theorists who work to either descriptively explain or prescriptively design institutions that achieve jointly superior in practices that failed to achieve mutual benefit and may be riddled with corruption.

Lewis surmises that the two critical facets of sociability, governance and language, do require moral obligation to comply with general conventions even when not directly in individuals' self-interest. The question is, what is the source of this normativity? Specifically, can it be deduced directly from simply preferring to live in civil society over a failed state, or a linguistic community over Babel? It seems clear from the fate of Rawls'

[51] John Rawls, 'Justice as Fairness,' in *John Rawls: Collected Papers*, ed. Samuel Freeman, Harvard University Press (Cambridge, 1999), 47–72, 60.

[52] See especially David Gauthier, *Morals by Agreement*, Clarendon Press (Oxford, 1986); Russell Hardin, *Morality within the Limits of Reason*, University of Chicago Press (Chicago, 1988), Martin Hollis, *Trust within Reason*, and Ken Binmore, *Game Theory and the Social Contract*, 2 vols., MIT Press (Cambridge, 1994–98).

[53] Rawls, 'Justice as Fairness,' 61.

argument sustaining fair play that deducing a momentary and case-by-case allegiance to rules of conduct, simply because an agent prefers that the rules hold in general rather than they do not hold at all, is insufficient within the confines of rational choice theory. Rawls, of course, argued that fair play is reasonable and does not depend on a metaphysical commitment to duty. However, Lewis' *Convention* attempts to wholly account for the normativity providing the cohesion throughout human society in terms consistent with the instrumentalism of expected utility theory.

Conclusion

Whereas he is post-positivist in accepting the thinnest metaphysics consistent with naturalism, Lewis is best regarded as a quintessential analytic philosopher. In response to Quine's unification of analytic and synthetic claims, Lewis attempts to save a concept of pure analyticity pertaining to statements that are true in all possible worlds. Here I have focused on Lewis' development of conventionalism. His unique move is to replace agreement or rarified normativity with instrumental preference satisfaction. To achieve this he works in the new tradition of formal game theory which provides a purely analytic statement of instrumental rationality.

I have traced Lewis' attempt to provide an instrumental explanation of the social contract and linguistic truthfulness. In defending his analysis, Lewis fights an intense battle to argue that the obligation to comply with conventional standards of truthfulness and lawful conduct may be instrumentally derived from the preference to live in a civil society over a state of nature or Babel. Lewis is hard-pressed to explain how agents will be motivated by conventions in those occasional circumstances in which one's interests are better served by lying or violating the law. In an argument strategy similar to Rawls' defense of fair play, Lewis suggests that in the exceptional cases in which cost-benefit analysis calculates that cheating or lying better maximizes expected utility, that an individual's sole motive for complying with a social convention is moral obligation. If Lewis' effort to vindicate moral obligation derived from purely instrumental considerations is deemed legitimate, then not only is Rawls' fair play consistent with rational choice theory, but it would seem to follow that general worries about cheating and free riding are overblown. However, the rational choice tradition has decided otherwise.[54]

Pressing home the point that Lewis' concept of convention does not encompass exceptional deviation, Hardin quotes Lewis who states, "One thing we do *not* tolerate is a convention to which most people want there to be exceptions, however few exceptions they want."[55] Hardin's point is that in social contracts, the elemental challenge is that many may well seek opportunistic exceptions for themselves (168). Lewis' reconciliation of a convention with a social contract, on the basis of the voluntary adoption of moral obligation to comply, would be deemphasized by subsequent theorists. Instead it became standard to view the challenge of maintaining a social contract on the model of the antagonistic Prisoner's dilemma played indefinitely. In infinite play, this repeated Prisoner's dilemma stands as a supergame in which agents may seek exceptions for themselves in a single round, but would rather comply over time if the cost of deviation is lost utility. Lewis' failure to satisfactorily instrumentalize the normativity underlying the social contract left ample room for the next generation to follow his lead by extending the concept of convention beyond regularities in behavior directly expressed in repeating contexts to regular patterns of behavior that extend across multiple rounds of interaction.[56] Interestingly, the quest for instrumental normativity as the ground for sociability is redirected to a complex strategic calculation over multiple rounds of play and long time periods. Although the original spirit of Lewis' investigation remains vibrant, these repeating games tend to locate conventional activity in groups as small as two individuals which stand a long distance from the large scale social and linguistic conventions Lewis originally endeavored to explain.

Acknowledgement

I am grateful to Kevin Scharp and Ben Caplan for their comments on and discussion of this paper.

[54] Hardin, *Collective Action*.

[55] Hardin, *Collective Action*, 162, Lewis, *Convention*, 77.

[56] See Hardin, *Collective Action*, 162–72.