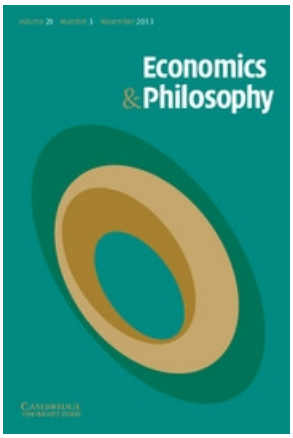

COMMON KNOWLEDGE, SALIENCE AND CONVENTION: A RECONSTRUCTION OF DAVID LEWIS' GAME THEORY

Robin P. Cubitt and Robert Sugden

Economics and Philosophy / Volume null / Issue 02 / October 2003, pp 175 - 210
DOI: 10.1017/S0266267103001123, Published online: 25 November 2003

Economics and Philosophy, **19** (2003) 175–210
DOI: 10.1017/S0266267103001123

COMMON KNOWLEDGE, SALIENCE AND CONVENTION: A RECONSTRUCTION OF DAVID LEWIS' GAME THEORY

Robin P. Cubitt and Robert Sugden
University of East Anglia

David Lewis is widely credited with the first formulation of common knowledge and the first rigorous analysis of convention. However, common knowledge and convention entered mainstream game theory only when they were formulated, later and independently, by other theorists. As a result, some of the most distinctive and valuable features of Lewis' game theory have been overlooked. We re-examine this theory by reconstructing key parts in a more formal way, extending it, and showing how it differs from more recent game theory. In contrast to current theories of common knowledge, Lewis' theory is based on an explicit analysis of the modes of reasoning that are accessible to rational individuals and so can be used to analyse the genesis of common knowledge. Lewis' analysis of convention emphasises the role of inductive reasoning and of salience in the maintenance of conventions over time.

INTRODUCTION

David Lewis is generally given credit for two major innovations in game theory: the first formulation of the concept of common knowledge and the first rigorous analysis of convention. Both of these innovations appear in

Earlier versions of this paper were presented at the 13th Amsterdam Colloquium at the University of Amsterdam, at a workshop on social norms at Wissenschaftskolleg zu Berlin, and at seminars at Tilburg University and the University of Bristol. We are grateful for comments from participants at those meetings, from two anonymous referees, and from Michael Bacharach, Nick Bardsley, Cristina Bicchieri, Luc Bovens, Simon Grant, David McCarthy, Shepley Orr, Brian Skyrms, Peter Vanderschraaf, Peter Wakker and Jörgen Weibull. Robert Sugden's work was supported by the Leverhulme Trust.

Convention: A Philosophical Study – the book, published in 1969, which grew out of Lewis' doctoral thesis. As the title suggests, this is primarily a work of philosophy. It is not addressed to game theorists, and few present-day game theorists appear to have read it in any detail. Common knowledge and convention entered mainstream game theory only when they were developed, later and independently of Lewis, by other theorists. Thus, although Lewis is usually credited with *priority*, his work has had relatively little *influence* on later developments. The prevailing view among game theorists seems to be that, although Lewis was brilliantly ahead of his time in 1969, his work has now been superseded. We shall argue that this judgement is mistaken.

Within a few years of the publication of *Convention*, the concept of common knowledge had been recognised as fundamental to game theory. Robert Aumann's (1976) analysis of common knowledge quickly came to be regarded as canonical. In the folk history of game theory, Lewis is often represented as the first person to think of the idea of common knowledge, but Aumann is almost universally credited with the first rigorous theoretical formulation.[1]

Lewis' analysis of convention is, as far as we know, the first formal analysis of games that are played recurrently in a population.[2] Until the late 1980s, game theory as generally understood by social scientists and mathematicians was the analysis of self-contained strategic interactions between ideally rational agents.[3] There was a presumption that an ideal theory would prescribe a unique rational strategy for each player in every possible game, identifiable deductively from features of the game itself. It is only relatively recently that game theorists have again considered the recurrent play of games within populations, and have countenanced the idea that each person's expectations about how current opponents will play may depend on what other opponents have been seen to do in the past. With this shift in the focus of game theory, convention has become a central concept. But this has happened as part of a more general movement towards evolutionary modes of analysis, and away from the assumption

[1] In fact, it may be that neither of these components of the folk history is strictly correct. Nozick (2001, p. 375, fn. 60) attributes the first formal statement of infinite layering of knowledge in game theory to his doctoral dissertation (Nozick, 1963). Another seminal work in the analysis of common knowledge is that of Schiffer (1972), whose approach is developed by Bacharach (1992).

[2] The idea that Nash equilibrium might be *interpreted* as a rest point in the dynamics of a game played recurrently by individuals drawn from a large population was suggested by Nash in his doctoral thesis (Nash, 1950, pp. 21–3; see Ritzberger and Weibull, 1995, pp. 1371–2). However, Lewis goes beyond this by developing an *analysis* of recurrent play.

[3] Games played repeatedly by *the same* individuals were analysed, but usually only by treating the whole series of 'stage games' played by those individuals as a single self-contained game.

of ideal rationality. This movement has drawn inspiration from the work of theoretical biologists. It has downplayed the role of reasoning – ideal or otherwise – in determining individual behaviour, focusing instead on blind or adaptive mechanisms of selection.[4] Lewis' theory of convention has been ignored or dismissed as depending on assumptions of rationality and common knowledge that are now thought to be redundant.

Our object in this paper is to re-examine Lewis' game theory, not as an episode in the history of ideas, but as a potential contribution to current theoretical analysis in social science. We introduce the main philosophical objectives of *Convention* in Section 1. Thereafter, we reconstruct key parts of Lewis' analysis in a more formal way, extend it, and show how it differs from more recent game theory. In Sections 3–5, we present a formal Lewisian analysis of common knowledge. In Sections 6–8, we examine Lewis' account of convention, drawing on the earlier analysis of common knowledge.

We shall argue that, far from having been superseded, Lewis' theory contains important ideas which have not been well understood, and whose explanatory power has yet to be exploited. We shall develop this argument in relation to the views of two philosophers of decision theory, Brian Skyrms and Peter Vanderschraaf, who have recently discussed Lewis' work. Skyrms (1996), whose argument we introduce in Section 2, compares Lewis' theory of convention with an evolutionary game-theoretic account, and claims that the evolutionary theory resolves problems unanswered by Lewis. Vanderschraaf (1998b) offers a reconstruction of Lewis' analysis of common knowledge which represents it within a theoretical framework that, in key respects, is similar to Aumann's. We indicate the main differences between these frameworks and Lewis' in Section 3, and expand on them in an appendix. In the paper as a whole, we argue that both Skyrms and Vanderschraaf fail to take account of some distinctive and valuable features of Lewis' theory.

1. LEWIS' ACCOUNT OF CONVENTION AND LANGUAGE

Lewis locates his analysis of convention in a philosophical tradition that derives from the work of David Hume.[5] He introduces it as 'a theory along

[4] During the 1970s and 1980s, an evolutionary form of game theory was developed by biologists, pioneered by Maynard Smith and Price (1973). This work was not influenced by Lewis, and was not much noticed by social scientists until the 1990s. As far as we know, its implications for economics and philosophy were first explored by Sugden (1986); among the first formal evolutionary game-theoretic models of convention was that of Young (1993). For overviews of evolutionary game theory, see Weibull (1995) and Samuelson (1997).

[5] In this respect, too, Lewis pioneers a path subsequently taken by other theorists. The idea that Hume's analysis of convention is game theory *avant la lettre* is now widely accepted (e.g. Sugden, 1986; Binmore, 1998; Vanderschraaf, 1998a).

the lines of Hume's, in his discussion of the origin of justice and property' (1969, p. 3). He endorses Hume's claim that convention:

> is only a general sense of common interest; which sense all the members of the society express to one another, and which induces them to regulate their conduct by certain rules. I observe, that it will be for my interest to leave another in the possession of his goods [i.e. to observe the convention of property], *provided* he will act in the same manner with regard to me. He is sensible of a like interest in the regulation of his conduct. When this common sense of interest is mutually express'd, and is known to both, it produces a suitable resolution and behaviour. And this may properly enough be call'd a convention or agreement betwixt us, tho' without the interposition of a promise; since the actions of each of us have a reference to those of the other, and are perform'd upon the supposition, that something is to be perform'd on the other part. (Hume, 1740/1978, p. 490)

From this passage, Hume goes on to say that the convention of property is an unintended consequence of repeated interactions: it emerges gradually, acquiring force as each person learns by experience that others can be relied on to follow it.

In an apparently offhand manner, Hume then remarks: 'In like manner are languages gradually establish'd by human conventions without any promise' (1740/1978, p. 490). Although Lewis does not explicitly discuss this one-sentence theory of language, the main philosophical objective of *Convention* is to formulate and defend a Humean account of language as convention. Lewis' aim is to show, contrary to the arguments of Bertrand Russell (1921), Willard Van Orman Quine (1936) and Morton White (1950), that there are conventions of language. Lewis attributes to these opponents the view that conventions, properly so called, must be created by agreement, and that agreement is not possible without the use of a pre-existing language. He claims to show that the formation of conventions need not involve the use of 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. (pp. 87–8)

Lewis goes on to explain how a certain kind of convention – a *signalling convention* – constitutes a 'rudimentary language' which conveys *meaning* (pp. 122–59).

In the present paper, our concern is with conventions and common knowledge, and with how these concepts should be represented in game theory. Because Lewis' analysis of convention is presented in support of a thesis about language, and because Skyrms' discussion starts from the question of whether Lewis' analysis adequately supports that thesis, we

will continue to touch on issues in the philosophy of language. But these are not in themselves our primary concern.

Lewis' argument focuses on a particular class of interactions between individuals, which can be defined in the following way. A (finite, non-cooperative) *game* is an interaction between two or more *players*; the number of players is finite and is denoted by m. Each player i chooses one (pure) *strategy* from a finite set S_i of possible strategies. For each *strategy profile* – that is, a list of m strategies, one for each player – there is a *payoff* to each player. For Lewis' purposes, it is sufficient to say that payoffs are real numbers, called *utility* indices, and that each player's payoff is a measure of the desirability to him of whatever is the outcome of the relevant strategy profile. Practical reason is interpreted in terms of desire and belief, so there is a presumption that each player aims for higher rather than lower payoffs for himself.[6]

A *coordination problem* is a special kind of game. As a first step in defining this concept, we introduce Lewis' concept of a 'proper coordination equilibrium' (we will simply say 'coordination equilibrium'). His definition is equivalent to this: for any given game, a *coordination equilibrium* is a strategy profile which satisfies the following two conditions. First, for each player i, i's strategy is strictly utility-maximising for him, given the strategies of the others (that is, the strategy profile is a strict Nash equilibrium). Second, for every pair of distinct players i and j, i's strategy is weakly utility-maximising for j, given the strategies of all players other than i. Thus, if any player makes a unilateral deviation from a coordination equilibrium, *no one* benefits from that deviation. A coordination problem is a game which has at least two coordination equilibria, and in which 'coincidence of interest predominates'. We take it from Lewis' examples that the intuitive idea is that all the players have a common interest in arriving at some coordination equilibrium, but do not much mind which of these equilibria they reach; their main problem is to make sure that they all head for the same equilibrium (pp. 5–24). It will turn out that, in relation to those parts of Lewis' theory that we reconstruct in this paper, the only significant property of a coordination problem is that it is a game with at least two strict Nash equilibria. (The other properties are relevant for those other parts of Lewis' theory that concern the relationship between conventions and norms.)

At the time Lewis wrote *Convention*, game-theoretic analysis was generally restricted to self-contained games. As we noted in the introduction, one of the original features of *Convention* is that it analyses

[6] Lewis does not make any strong claims about utility or preference. He interprets utility indices as no more than 'rough indications of strength of preference', and assumes only that 'decision theory applies in some approximate way to ordinary rational agents with imperfectly coherent preferences' (p. 9).

games that are played recurrently within a population. The idea seems to be that each instance of a game involves m players drawn from a population which may be larger than m; different interactions may involve different combinations of m individuals as players of the game. We say 'seems' because Lewis does not give an explicit model of the process by which players are selected. However, he defines 'convention' in terms of 'the behaviour of members of a population P when they are agents in a recurrent situation'; the situation that recurs is modelled as a game (p. 42). In some of Lewis' examples of conventions, the 'population' is simply a fixed set of m players of a given coordination problem. One such example concerns two people who meet every week; another – borrowed from Hume – involves the behaviour over time of the two rowers of a boat. But, in other examples, the population is clearly much larger than m. One of these is a rule once followed by phone users in Lewis' home town of Oberlin, Ohio, prescribing that when a call was cut off (as they routinely were), the original caller called back: here $m = 2$ but the population contains all Oberlin phone users (pp. 42–4).

Lewis' analysis applies to coordination problems that are recurrent situations within populations. His definition of a convention is equivalent to the following: a *convention* is a regularity of behaviour in such a game within such a population, such that it is both true and common knowledge (i) that all members of the population, when playing the game, choose to play their parts of some particular coordination equilibrium and (ii) that everyone expects everyone else to conform to this regularity (p. 58).[7] We postpone until Section 3 the question of what Lewis means by 'common knowledge'.

Lewis' analysis of language is conducted in terms of a special kind of coordination problem: a *signalling problem*. This is a coordination problem involving two players, the *audience* (A) and the *communicator* (B). B, but not A, observes which of a set of possible states of nature is the case and then chooses which of a set of actions to perform. Having observed B's action, A chooses from a set of actions. A strategy is interpreted as a complete specification of what a player will do in every contingency, and payoffs are expressed in units of *expected* utility. This idea is most easily explained by example.

Consider the following Reversing Game (adapted from one of Lewis' examples). A is a truck driver reversing into a constricted space; B is an assistant who shares A's desire that the truck is reversed successfully. B is

[7] Strictly, this is a preliminary definition. Lewis' final definition allows for some fuzziness: a regularity can count as a convention if in *almost all* instances of the recurrent situation the interaction is a coordination problem, if *almost all* players *almost always* conform to the regularity, and so on (p. 78). In this paper, we work with the definition that we give in the main text.

visible to A and, unlike A, observes whether or not there is room for the truck to reverse further. Conditional on this observation, B chooses from a set of alternative gestures. A observes B's action; conditional on this observation, she chooses whether the truck moves or stops. An example of a strategy for B is: 'If room to reverse, make beckoning gesture; if not, show hands with palms outwards'. An example of a strategy for A is: 'If B makes beckoning gesture, reverse; if not, stop'. In the simplest possible case, there are just two possible states of nature to be observed by B, two possible gestures for B, and two possible actions for A. Since each player can condition either possible action on either possible observation, this gives a coordination problem in which there are four alternative pure strategies for each player and two coordination equilibria. In each of these equilibria, B makes a particular gesture if and only if he observes that there is room to reverse, and A reverses if and only if she observes that gesture. These equilibria differ in terms of *which* gesture is given when there is room to reverse.

Now consider a signalling problem that is faced repeatedly by the members of some population. According to Lewis, a convention for such a problem – a *signalling convention* – is a simple language.[8] The signals used in such a convention have *meaning* by virtue of their reliable association with particular states of the world, intentions and actions, and by virtue of the fact that these associations are common knowledge (pp. 143–59). For example, suppose that in the Reversing Game, it is true and common knowledge: (i) that truck drivers reverse in response to (and only in response to) beckoning gestures from their assistants, and that assistants make beckoning gestures if (and only if) there is room for the truck to reverse; and (ii) that everyone expects everyone else to conform to this regularity. Then, on Lewis' account, the beckoning gesture *means* 'There is room to reverse' or 'Reverse!', and this meaning is established by convention. If we accept Lewis' analysis of convention, and if we accept his claim that signalling conventions are rudimentary languages, Lewis has discharged the task he set himself: he has shown that there can be conventions of language.

2. SKYRMS' CRITIQUE

Has Lewis really met the challenge set by Quine and his fellow sceptics? Skyrms (1996) argues that the answer is 'No'. Nevertheless, Skyrms also locates himself in the Humean tradition, and endorses Hume's account of

[8] Lewis (pp. 141–3, 160) chooses to reserve the term 'language' for signalling conventions which use either vocalisations or inscriptions. This restriction seems arbitrary, particularly in view of the role played by hand gestures in all spoken human languages.

convention.[9] He is convinced that the Quinean challenge can be met by an evolutionary analysis of convention. Although Skyrms acknowledges Lewis' achievements as a game theorist, the main thrust of his argument is that Lewis' analysis of convention has been superseded by evolutionary game theory. In responding to Skyrms, we shall argue that evolutionary game theory can still learn from Lewis. While remaining agnostic on the question of whether either Lewis or Skyrms ultimately meets the sceptical challenge about language, we shall maintain that Skyrms' principal criticisms of Lewis' analysis of convention are misplaced.

Skyrms' objections focus on two features of Lewis' game-theoretic reasoning. First, according to Skyrms, Lewis *assumes* that the structure of each coordination problem and the strategies chosen by the players are common knowledge. Second, Lewis' analysis of how players coordinate their strategies depends on the idea, originally due to Thomas Schelling (1960), that certain equilibria have non-rational properties of *salience* which attract the attention of all players. Skyrms (pp. 83–4) summarises Lewis' argument by saying that the assumption of salience is used to explain how one particular convention is first selected from the set of possible conventions, and that the assumption of common knowledge is used to explain why, once any convention has been selected, no one deviates from it. Later, we shall argue for a different reading of Lewis.

According to Skyrms (p. 84), a Quinean sceptic is entitled to ask the questions: '*Where does all the common knowledge come from?*' and '*Where is the salient equilibrium?*' Skyrms offers the sceptic the counter-argument that 'an explanation of the amount of common knowledge assumed [by Lewis] might require far more pre-existing communication than is explained by the game under consideration'. Thus, Lewis' argument might rely on smuggled-in assumptions about mutual understanding, and language might be needed to explain how this understanding arises. The assumption of salience might be vulnerable to the same objection. So, Skyrms says, the sceptic can still charge Lewis with circularity.[10] Further, Skyrms is sceptical about whether salience exists at all in some signalling games. In the model signalling games analysed by Lewis and by Skyrms himself, the alternative conventions are (according to Skyrms) completely symmetric, and so no signalling system can be uniquely salient. We take it that Skyrms thinks the same is true of some of the real-world situations

[9] Skyrms (p. vi) takes as an epigram for his book a quotation from the passage in which Hume uses language as an example of a convention, and in which he argues that conventions arise gradually and acquire force by a slow progression. Although Skyrms does not comment on this quotation, we assume that he endorses it.

[10] Similar criticisms of Lewis' theory can be found in Skyrms (1990, pp. 52–4) and Vanderschraaf (1995, pp. 81–3).

in which signalling systems are known to have evolved (pp. 84, 92–3, 102).[11,12]

Skyrms claims to improve on Lewis' argument by dropping the assumptions of common knowledge and salience. Instead, he uses the methods of evolutionary game theory. For Skyrms, this means borrowing theoretical tools that were designed for modelling biological evolution, and using them, in place of rationality and common knowledge, to explain processes of cultural evolution. Skyrms relies on random variation to create asymmetries in the frequencies with which different strategies are played in the population. In a recurrent signalling problem, the strategies that are played more frequently are more successful, and so have a greater tendency to replicate; in consequence, initially random asymmetries are amplified by differential replication. By virtue of this mechanism, the emergence of *some* signalling system is a 'moral certainty'. However, *which* system emerges can be sensitive to small differences in the initial values of crucial parameters; in Skyrms' model, this is 'a matter of chance, not salience' (p. 93).

3. LEWIS' THEORY OF COMMON KNOWLEDGE

We begin by looking at how Lewis answers Skyrms' first question: Where does all the common knowledge come from? To a reader who is familiar with modern game theory, this question might suggest that Lewis assumes the now-standard conception of infinitely iterated knowledge (that is, person i knows that person j knows that ... that person k knows x) and that Skyrms is questioning the justification for doing so. But to understand Lewis' concept of common knowledge and its role in his analysis of convention, it is essential to realise that his definition of common knowledge is *not* an informal version either of the standard iterated knowledge assumption or of Aumann's now-canonical model of that.

If Lewis had invoked Aumann's model, it would have been entirely appropriate to ask where all the common knowledge comes from. As Aumann (1987) explains, his model describes a universe in which everything, apart from which state of the world is the true state, is, in

[11] The claim that certain *model* games lack salient coordination equilibria does little by itself to support the argument that we can explain the evolution of signalling conventions *in the real world* without appealing to salience. If the lack of salience in these models is the result of modelling assumptions, we can always ask whether those assumptions adequately represent that subset of real signalling problems for which signalling conventions have in fact evolved.

[12] In a later paper, Skyrms (1999) acknowledges that some signalling conventions might have originated from 'natural salience', but, except where there is direct evidence to support such a hypothesis, he remains sceptical of explanations of conventions which appeal to salience.

an informal sense, common knowledge. Within the model, 'knowledge' and 'common knowledge' have specific theoretical meanings; in order to interpret these concepts, it is necessary to *assume* common knowledge in the informal sense. Clearly, a model of this kind is not appropriate for an analysis of the genesis of common knowledge. But Aumann's approach is not Lewis'. In Appendix 2, we describe Aumann's theory of common knowledge, and show how radically it differs from Lewis'. We also show how these differences are (in our view unhelpfully) edited out in Vanderschraaf's (1998b) reconstruction of Lewis' theory.

One distinctive feature of Lewis' analysis is that it is concerned with what a person has *reason to believe*. If warranted *true* belief is a requirement of 'knowledge', then Lewis' analysis is not, strictly speaking, about knowledge at all; it is about warranted belief. The concern is with those modes of human reasoning, whether deductive or inductive, that can properly be said to justify beliefs or actions. A belief might be justified according to reasonable standards of inductive inference, yet not be true.

Although it is an essential part of Lewis' theory that human beings are *to some degree* rational, he does not want to make the strong rationality assumptions of conventional decision theory or game theory. He distinguishes between what an individual *has reason to believe* and what she *actually believes*.

Our interpretation of 'reason to believe' is as follows. To say that some individual i has reason to believe some proposition x is to say that x is true within some logic of reasoning that is *endorsed* by (that is, accepted as a normative standard by) person i. For x to be true within such a logic of reasoning, it must *either* be treated as self-evident *or* be derivable from propositions that are treated as self-evident using the inference rules of the logic. Self-evidence may be either *a priori* or obtained through observation; the rules of inference may be deductive or inductive.

To say that i actually believes x is to state a proposition about i's psychological state. For example, anyone who accepts the rules of arithmetic has reason to believe that $618 \times 377 = 232{,}986$, but most of us, most of the time, do not hold any firm beliefs about the truth or falsity of that proposition. Lewis presents a theory of practical reasoning which imposes strong consistency conditions on what any person can have reason to believe, but he does not assume that people always believe what they have reason to believe. However, he maintains that there is enough connection between reason to believe and actual belief for the former to be useful in explaining human behaviour (p. 141).

Before presenting Lewis' definition of common knowledge, we need some notation and some other definitions. For the most part, the notation is ours rather than Lewis'. We use i, j, k and l to denote persons, P to denote a population (i.e. a set of persons), and x, y and z to denote propositions. For any person i and any proposition x, $R_i(x)$ denotes 'i has reason to believe

that x'. Such formulae can be nested to any finite depth; for example, $R_i[R_j(x)]$ denotes 'i has reason to believe that j has reason to believe that x'. For any population P and any proposition x, we use the notation $r^P(x)$ to denote that all finitely nested formulae $R_i[R_j[\ldots[R_k(x)]\ldots]]$ are true for all i, j, . . . , k in P. That is, $r^P(x)$ is true if and only if, for all persons i, j, k, . . . in P: i has reason to believe that x, i has reason to believe that j has reason to believe that x, i has reason to believe that j has reason to believe that k has reason to believe that x, and so on.

Readers who know of Lewis' analysis only through the folk history of game theory may now expect us to say that $r^P(x)$ is Lewis' definition of 'x is common knowledge in P'. But it is not.[13] Lewis presents a set of conditions which he shows are sufficient to make $r^P(x)$ true; *that set of conditions* is his definition of common knowledge. Thus, Lewis offers an *analysis* of how, given a certain state of affairs, $r^P(x)$ can come to hold for some proposition x and some population P. If $r^P(x)$ holds for some P and x, we shall say that in P there is *iterated reason to believe* that x.

Given this formulation of Lewis' system, Skyrms' first question can be rephrased as: Where does all the iterated reason to believe come from? The formal component of Lewis' answer is provided by the set of conditions he presents and by the proof that those conditions imply $r^P(x)$. In the case of convention, the relevant x is the continuation of some regularity of behaviour that is consistent with a coordination equilibrium. In the remainder of this section and in Section 4, we state conditions under which iterated reason to believe in an arbitrary proposition x is generated. In Sections 6 and 7, in discussing Lewis' account of the reproduction of conventions, we consider circumstances under which the formal conditions can and cannot be met.

For Lewis, what any person i has reason to believe may depend on i's 'background information', which is not necessarily the same as anyone else's background information. It may also depend on the 'inductive standards' that i uses, and these may be different from the inductive standards used by others. However, as we shall explain later, common knowledge in Lewis' sense is possible only when individuals have reason to believe that, in particular relevant respects, they have common background information and common inductive standards.

Lewis implicitly assumes that each person's inductive standards are reasonable. He does not model these standards explicitly, but relies on his readers' intuitive sense of the meanings of the formulae with which he describes them. This can make his arguments difficult to follow. We

[13] That Lewis does not use $r^P(x)$ as the definition of common knowledge is also pointed out by Vanderschraaf (1998b), as the starting point for his own reconstruction of Lewis' analysis.

prefer to be more explicit, by stating as postulates those formal properties of Lewis' logical operators that are necessary for our proofs.

In addition to propositions, Lewis treats *states of affairs* as primitive. States of affairs (which we will denote by A and A′) are alternative specifications of how the world, as seen by the modeller, really might be. Lewis does not impose any particular structure on states of affairs. One possible interpretation is that they correspond roughly with 'states of the world' in Leonard Savage's (1954) subjective expected utility theory. The assertion that a particular state of affairs A is in fact the case (which we write as 'A holds') is a proposition. The distinction between the state of affairs A and the proposition 'A holds' does not do a great deal of work in Lewis' theory. However, it is useful in giving us, as modellers, a language in which we can *refer* to alternative ways the world might be, without *asserting* anything about those entities.[14] Individuals within Lewis' framework have beliefs about, and have reason to have beliefs about, propositions. His theory is primarily about what individuals have reason to believe, given the states of affairs that actually hold.

At the heart of Lewis' theory is a three-place relation of *indication*, which governs individuals' inferences from states of affairs that are believed to hold to propositions that are believed to be true. Lewis (pp. 52–3) defines the formula 'A indicates to i that x', which we write as 'A ind_i x', as '*if* i has reason to believe that A holds, i *thereby* has reason to believe that x is true'. He does not spell out the exact logical status of the *if* ... *thereby* ... formula. However, by using 'i thereby has reason to believe' instead of 'then i has reason to believe', Lewis clearly intends *if* ... *thereby* ... to be stronger than the material implication, $\Rightarrow$. On the most natural reading of the definition of 'A ind_i x', i's reason to believe that A holds *provides i's reason for* believing that x is true. The definition implies the following property of the indication relation:[15]

(A1) For all persons i, for all states of affairs A, for all propositions x: $[R_i(\text{A holds}) \wedge (\text{A ind}_i\ \text{x})] \Rightarrow R_i(\text{x})$.

That is, if i has reason to believe that A holds, and if that reason provides i with reason for believing x, then i also has reason to believe x.

[14] It is not required that *every* proposition assert that some state of affairs holds. This allows the user of Lewis' theory some freedom to specify, for any particular model, what the set of alternative states of affairs is and what structure it should have.

[15] A1 is equivalent to $[\text{A ind}_i\ \text{x}] \Rightarrow [R_i(\text{A holds}) \Rightarrow R_i(\text{x})]$. However, A ind_i x is *not* equivalent to $[R_i(\text{A holds}) \Rightarrow R_i(\text{x})]$. The material implication $[R_i(\text{A holds}) \Rightarrow R_i(\text{x})]$ states that, if in fact it is the case that i has reason to believe that A holds, then it is also the case that he has reason to believe x. But it does not follow from this that i's reason to believe x has any relation to his reason to believe that A holds. This last step, however, *is* required for indication.

Our interpretation of the formula 'A ind_i x' is that, in the logic of reasoning that i endorses, there is an inference rule which legitimates inferring x from 'A holds'.[16] The set of inference rules endorsed by an individual will include principles of inductive inference. However, it seems clear that Lewis also intends that the inference rules of each person's logic include the standard rules of deductive inference. At an intuitive level, this idea is straightforward enough, but representing it formally is more difficult. The problem is that the concept of a person's logic of reasoning never appears explicitly in Lewis' analysis. His analysis is conducted in terms of a non-standard logical relation, indication, the formal properties of which are not self-evident.

For the purposes of this paper, we do not need to specify *all* the properties of indication implied by the assumption that, within each person's logic, deductive inferences are legitimate. We simply state the five such properties which will be used in our proofs. Of these, one is used in the proof of Lewis' main conclusion. The others are used to extend his analysis in particular ways. We stress that these are *not* independently-motivated axioms: given our interpretations of 'reason to believe' and 'indication', they are all implications of Lewis' assumption that each person endorses deductive inference.

These properties are:

(A2) For all persons i, for all states of affairs A, A′: [(A holds) entails (A′ holds)] $\Rightarrow$ A ind_i (A′ holds).
(A3) For all persons i, for all states of affairs A, for all propositions x, y: [(A ind_i x) $\wedge$ (A ind_i y)] $\Rightarrow$ A ind_i (x $\wedge$ y).
(A4) For all persons i, for all states of affairs A, A′, for all propositions x: [(A ind_i [A′ holds]) $\wedge$ (A′ ind_i x)] $\Rightarrow$ A ind_i x.
(A5) For all persons i, for all states of affairs A, for all propositions x, y: [(A ind_i x) $\wedge$ (x entails y)] $\Rightarrow$ A ind_i y.
(A6) For all persons i, j, for all states of affairs A, A′, for all propositions x: [(A ind_i R_j[A′ holds]) $\wedge$ R_i(A′ ind_j x)] $\Rightarrow$ A ind_i R_j(x).

If one proposition entails another, the second can be inferred from the first: hence A2. If, from some proposition, each of two propositions x and

[16] In Lewis' model, subjective probabilities are not used. Thus, if i 'has reason to believe that' x is true, i reasons about x as if its truth was a matter of subjective certainty. In a general model of inductive reasoning, this could lead to problems: there might be conceivable (although improbable) states of affairs in which two independent and normally reliable rules of inductive inference have contradictory implications. To avoid this problem, propositions arrived at by inductive inference must be treated as in some way provisional or corrigible. Since Lewis does not discuss this problem, and since it is not significant for what follows, we note it and pass on.

y can be inferred, then x ∧ y can be inferred from that first proposition: hence A3. If a second proposition can be inferred from a first, and if a third proposition can be inferred from the second, then the third can be inferred from the first: hence A4 and A5. A6 says that if i's logic legitimates an inference from 'A holds' to the proposition that j has reason to believe 'A′ holds', and if i has reason to believe that j's logic legitimates an inference from 'A′ holds' to x, then i's logic legitimates an inference from 'A holds' to the proposition that j has reason to believe x.[17]

We are now in a position to state the definitions that are central to our reconstruction of Lewis' theory.

In any given population P, a state of affairs A is a *reflexive common indicator that* x if, and only if, the following four conditions hold:[18]

(C1) For all persons i in P: A holds $\Rightarrow R_i$(A holds).
(C2) For all persons i, j in P: A ind_i R_j(A holds).
(C3) For all persons i in P: A ind_i x.
(C4) For all persons i, j in P, for all propositions y: (A ind_i y) $\Rightarrow R_i$[A ind_j y].

We have not asserted that C1–C4 hold for *all* A and x, or even, yet, for *any* A and x; they merely constitute a definition of a reflexive common indicator. A state of affairs A which has the properties C1 and C2 is, in particular senses, *self-revealing* and *public*: if in fact A holds, then everyone has reason to believe that A holds, and anyone who has reason to believe that A holds thereby has reason to believe that everyone has reason to believe that A holds. If C3 holds for some A and x, each person's logic of reasoning allows an inference from 'A holds' to x. C1, C2 and C3 are stated formally by Lewis (subject to the proviso noted in fn. 18); we have merely translated them into our notation. In contrast, C4 is a property that Lewis states only as 'suitable ancillary premises regarding our rationality,

[17] If i has reason to believe that j's logic legitimates an inference from (A′ holds) to x, then, because *i's* logic obeys the rules of deductive inference, i has reason to believe [R_j(A′ holds) $\Rightarrow R_j$(x)]. Hence, *given the antecedent of the previous sentence*, if i's logic allows an inference from (A holds) to R_j(A′ holds), it also allows an inference from (A holds) to R_j(x). Notice that A6 attributes deductive inference only to i's reasons to believe. It does not postulate anything about what i has reason to believe about the inference rules endorsed by j.

[18] Our definition corresponds with Lewis' definition of a *basis for common knowledge*, except that, read literally, Lewis' version of C1 is: for all persons i in P: R_i(A holds). We suggest that Lewis' intention is better represented by C1. We use a new term for Lewis' concept of a basis for common knowledge to avoid confusion with other definitions of common knowledge. We use the word 'reflexive' for the following reason: if some state of affairs A is a reflexive common indicator in P that some proposition x is true, it is also a reflexive common indicator in P that 'A holds' is true. (This follows from the trivial formula A ind_i [A holds], which is licensed by A2.)

inductive standards, and background information'. In an example in which the population consists of 'you' and 'I', he fleshes out these premises as: 'Suppose you and I do have reason to believe we share the same inductive standards and background information, at least nearly enough so that A will indicate the same things to both of us' (p. 53). C4 formalises that supposition: if A indicates any particular proposition y to any person i, then i has reason to believe that A indicates y to any person j.[19]

To show how C1–C4 might be satisfied, consider an example. Suppose that P contains just two normally sighted individuals, i and j. Let A be the state of affairs that i and j are together in the same room when that room is lit by a flash of lightning. Let x be the proposition 'within a few seconds, there will be the noise of thunder'. The nature of A is such that C1 and C2 can be presumed to hold: if in fact there has been a flash of lightning in such circumstances, each person in the room has reason to believe that there has been a flash of lightning, and thereby that the other person has reason to believe that there has been a flash of lightning. We can take it that 'from (A holds), infer x' (roughly: from lightning, infer thunder) is an inductive inference that is endorsed by all normal adults. That is, A indicates to both i and j that x is true, yielding C3. So, provided that each person has reason to believe that the other shares his own inductive standards and background information about what can be inferred from lightning (that is, that C4 is satisfied), A is a reflexive common indicator in P that x.

Now suppose that A holds: it is in fact the case that i and j are together when a flash of lightning occurs. What does this imply about what i and j have reason to believe? The core of Lewis' analysis is distilled in the following theorem, which is proved in Appendix 1:

Lewis' Theorem: For all states of affairs A, for all propositions x, and for all populations P: if A holds, and if A is a reflexive common indicator in P that x, then $r^P(x)$ is true.

Lewis' definition of common knowledge is equivalent to the following: a proposition x is *common knowledge* in a population P if and only if some state of affairs A holds, such that A is a reflexive common indicator in P that x. By virtue of Lewis' Theorem, 'x is common knowledge in P', defined in this way, *implies* $r^P(x)$. However, the converse implication is not generally true, and so Lewis' definition of common knowledge in P that x is *not* equivalent to $r^P(x)$.

Consider the following case. Suppose there is some person k, not herself a member of P, whose statements are treated as authoritative by

[19] C4 is a stronger condition than is strictly necessary. It will become clear from our proofs that we do not require the implication in C4 to hold for *all* propositions y, but only for certain types of proposition regarding first- and higher-order reasons to believe that A holds.

each member of P. That is, for each i in P, for any proposition z, the state of affairs 'k states to i that z is true' indicates to i that z. Now suppose that, separately and privately to each member of P, k states that both x and $r^P(x)$ are true. Each member of P has reason to believe that these statements have been made to him. Thus, we have $R_i(x)$ and $R_i[r^P(x)]$ for all i in P, which, given that reason to believe obeys the rules of deductive logic, imply that $r^P(x)$ is indeed true. However, there may be no state of affairs A such that A holds and A is a reflexive common indicator in P that x. For example, each member of P may have no reason to believe that k's statement has been made to the others. Or he may have no reason to believe that the others treat k's statements as authoritative.

4. DISTRIBUTED REASON TO BELIEVE

Lewis presents his theorem about common knowledge in the context of a coordination problem that is solved 'by agreement'. Two people, 'you' and 'I', are meeting today. We have a common interest in meeting again tomorrow. On leaving, you say to me that you will return to the same place tomorrow. Lewis uses A to represent the state of affairs just described, and x to represent the proposition that you will return. He then explains the conditions under which A is a reflexive common indicator that x is true, and hence (by Lewis' Theorem) under which there is iterated reason to believe x (pp. 52–6). In this case, iterated reason to believe that a coordination problem will be solved in a particular way is generated by a state of affairs which (like the lightning in our previous example) is self-revealing and public. For many conventions, however, there does not *seem* to be any such state of affairs to indicate that the convention will be followed.

Take a case which Lewis uses as one of his examples of convention: the convention that, in America, people drive on the right. Is there a self-revealing and public state of affairs which indicates that future coordination problems among American drivers will be resolved by their keeping right? The obvious candidate is the state of affairs that, up to now, almost all American drivers have kept right. But if this state of affairs satisfies all of C1–C4, it does so indirectly, by virtue of people's having reason to make inferences that are not represented explicitly in Lewis' formal model. As Lewis recognises, no American driver is directly aware of the driving habits of *all* Americans:

> I know very well that I have often seen cars driven in the United States, and almost always they were on the right. And since I have no reason to think I encountered an abnormal sample, I infer that drivers in the United States

> do almost always drive on the right; so anyone I meet driving in the United States will believe this just as I do, will expect me to believe it, and so on . . . . Given a regularity in past cases, we may reasonably extrapolate it into the (near) future. (p. 41)

Lewis does not go beyond this informal discussion. We now extend our formalisation of Lewis' analysis to show how iterated reason to believe can be generated by states of affairs which, in their entirety, do not directly reveal themselves to anyone.

Consider any population $P = \{1, \ldots, n\}$ and any state of affairs $A = (A_1$ *and* $\ldots$ *and* $A_n)$,[20] where each A_i is a state of affairs. Let x be any proposition. We shall say that A is a *distributed indicator in* P *that* x if and only if the following conditions hold:

(D1) For all persons i in P: A_i holds $\Rightarrow R_i(A_i$ holds).
(D2) For all persons i, j in P: A_j ind_i $R_j(A_j$ holds).
(D3) For all persons i, j in P: A_j ind_i (A holds).
(D4) For all persons i in P: A ind_i x.
(D5) For all persons i, j in P, for all propositions y, for $A' = A$ and for $A' = A_k$ for any k in P: $(A'$ ind_i y) $\Rightarrow R_i[A'$ $ind_j y]$.

Applying this analysis to the example of driving on the right, let P be the population of American drivers. For each person i, let A_i be the state of affairs that, up to today, almost all the American drivers that i has seen have driven on the right. Let $A = (A_1$ *and* $\ldots$ *and* $A_n)$, that is, A is the state of affairs that, up to today, almost all the American drivers that *anyone* has seen have driven on the right. Let x be the proposition that, tomorrow, almost all American drivers will drive on the right. To say that D1 holds is to say that if almost all the American drivers i has seen have driven on the right, then i has reason to believe that this is in fact the case. To say that D2 holds is to say that, for any two persons i and j, if i has reason to believe that all the American drivers that j has seen have driven on the right, i thereby has reason to believe that j has reason to believe that those drivers drove on the right. To say that D3 holds is to say that each person has reason to make an inductive inference from a regularity observed in a sample of observations of American driving (almost all the American drivers that a particular person has seen have kept right) to a regularity in a larger universe (almost all the American drivers *anyone* has seen have kept right). To say that D4 holds is to say that each person has reason to make an inductive inference from a regularity in the past (almost all

[20] If states of affairs are treated as equivalent to Savage events (which have a set-theoretic structure), $A = A_1 \cap \ldots \cap A_n$.

the American drivers anyone has seen up to now have kept right) to a regularity that is predicted to occur in the future (tomorrow, almost all American drivers will keep right). To say that D5 holds is to say that each person has reason to believe that, in relevant respects, other people share his own inductive standards and background information – that is, that they will infer the same conclusions from propositions about observed regularities in driving behaviour.[21]

The following theorem is proved in Appendix 1:

Distribution Theorem: For all propositions x and for all populations $P = \{1, \ldots, n\}$: if some $A = (A_1$ *and* $\ldots$ *and* $A_n)$ is a distributed indicator in P that x, then A is also a reflexive common indicator in P that x.

The following result follows immediately from the conjunction of the Distribution Theorem and Lewis' Theorem:

Corollary 1: For all populations P, for all states of affairs A, for all propositions x: if A holds, and if A is a distributed indicator in P that x, then $r^P(x)$ is true.

This provides an analysis of how, given that certain modes of reasoning are shared within a population P, the existence of a state of affairs $A = (A_1$ *and* $\ldots$ *and* $A_n)$ can induce iterated reason to believe the truth of a proposition x.

5. ACTUAL BELIEF

So far, we have said nothing about what individuals *actually* believe. Reasons to believe, not actual beliefs, are the focus of Lewis' analysis. As we shall show later, Lewis assumes remarkably little about the extent to which individuals actually believe what they have reason to believe. However, game theorists may wonder about the implications of the analysis for the special case in which individuals believe everything they have reason to believe. In this Section, we draw out those implications.

Let $b_i(x)$ denote 'i believes that x'. Let $b^P(x)$ denote that all finitely nested formulae $b_i[b_j[\ldots[b_k(x)]\ldots]]$ are true for i, j, ..., k in P. That is, $b^P(x)$ is true if and only if, for all persons i, j, k, ... in P: i believes that x, i believes that j believes that x, i believes that j believes that k believes that x, and so on. If $b^P(x)$ holds for some P and x, we shall say that in P there is *iterated actual belief* that x.

[21] Like C4, D5 is stronger than is strictly necessary: we do not require the implication to hold for all propositions y, but only for certain types of proposition, as will become clear in our proofs.

We shall say that a person i *reasons faultlessly* if, for all propositions x, $R_i(x)$ implies $b_i(x)$. In other words: a person who reasons faultlessly believes everything she has reason to believe.[22]

The following theorem is proved in Appendix 1:

Iterated Belief Theorem: Consider any state of affairs A, any proposition x, and any population P, such that (i) A is a reflexive common indicator in P that x and (ii) A is a reflexive common indicator in P that each person in P reasons faultlessly. Suppose that A holds and that each person in P reasons faultlessly. Then $b^P(x)$ is true.

Hence:

Corollary 2: Consider any state of affairs A and any population P such that A is a reflexive common indicator in P that each person in P reasons faultlessly. Suppose that A holds and each person in P reasons faultlessly. Then b^P(Each person in P reasons faultlessly) is true.

The proposition b^P(Each person in P reasons faultlessly) is the closest analogue in Lewis' system to the familiar game-theoretic assumption of infinitely iterated knowledge of players' rationality. As such, it is useful in relating Lewis' analysis of common knowledge to the analysis that is now conventional in game theory, and it suggests that Lewis' theoretical framework might prove useful in analysing how iterated belief in individuals' reasoning abilities could come about. However, in fact, Lewis does not assume that people reason faultlessly, nor that there is iterated reason to believe that they do so; we do not make these assumptions either. In the next section, we explain what is required of people's reasoning for a convention to exist.

6. HOW REGULARITIES GENERATE CONVENTIONS

From now on, we concern ourselves only with *tacit* conventions. That is, our concern is with cases in which each person's reasons for following a convention do not derive from prior communication with other members of the population (as in the case of an exchange of promises); they derive only from each person's previous experience of regularities in other people's behaviour. Recall that, for Lewis, a convention is a regularity of behaviour within a population, such that it is both true and common knowledge that everyone plays his part in some particular coordination

[22] We use the expression 'reasons faultlessly' rather than the more usual 'is rational' to signal that the requirement is that i reasons according to the standards that *she* endorses. As modellers, we are not asserting that these standards are uniquely correct.

equilibrium, and that everyone expects everyone else to conform to this regularity. Suppose that, up to some point in time, everyone has in fact conformed to some regularity that is a coordination equilibrium. What, if anything, makes that regularity a convention? That question can be decomposed into three. How does there come to be iterated reason to believe that the regularity will persist? How does it come about that everyone actually expects the regularity to persist? And how does it come about that the regularity in fact persists?

Consider again the example of American driving, introduced in Section 4. According to Lewis, the regularity that Americans drive on the right is a convention. What makes that claim true? Recall that $A = (A_1$ *and* ... *and* $A_n)$, where A_i is the state of affairs that, up to today, almost all the American drivers that i has seen have driven on the right, and that x is the proposition that, tomorrow, almost all American drivers will drive on the right. Clearly, A holds: it is a fact that, up to now, Americans have driven on the right, and so that almost all Americans that anyone has seen have kept right. But we must ask: Is there iterated reason to believe that x is true? Does everyone in fact believe that x is true? And is x true?

We start with the first question. By using Corollary 1, we can state five conditions which are jointly sufficient to ensure that, in P, there is iterated reason to believe that x is true. These conditions are D1–D5. In Section 4, we explained what these conditions mean in the case of driving on the right. D1 and D2 seem relatively undemanding, given the public visibility of driving behaviour. D3 and D4 require that each person's logic of reasoning allows certain kinds of inductive inference. D5 requires that, in relevant respects, each person's logic of reasoning attributes certain of its own standards of inference to other people's logics. In particular, this must be the case for the standards of inductive inference that make D3 and D4 true.

So a central feature of Lewisian analysis of tacit conventions is that it depends on assumptions about shared standards of inductive inference. It is in trying to justify these assumptions that Lewis invokes the idea of salience.

One of the distinctive features of Lewis' account of convention is its recognition that there is a problem in explaining how any coordination equilibrium, once reached in a population, tends to maintain itself. From the perspective of classical game theory, the puzzle is to explain why, for rational players of a fully-specified game, information about how other players have behaved in similar games in the past has any relevance at all. Why isn't the behaviour of the present players fully determined by the nature of the game they are now about to play? Why aren't bygones bygones? From the perspective of philosophy, this is an instance of the familiar problem of induction.

Lewis' answer arises out of his discussion of different ways in which one-off coordination problems might be solved. One possibility is a pre-play exchange of promises between the players. Another is a pre-play exchange of declarations of present intention, which fall short of promises. Introducing a third possibility, Lewis refers to Schelling's (1960) experiments, which show that one-off coordination problems can sometimes be solved without prior communication, in virtue of the players' recognition that one of the available coordination equilibria is *salient*, which Lewis glosses as 'one that stands out from the rest by its uniqueness in some conspicuous respect'. Lewis offers the following explanation of the reasoning that leads people to coordinate on salient equlilibria in Schelling's problems:[23]

> How can we explain coordination by salience? The subjects might all tend to pick the salient as a last resort, when they have no stronger ground for choice. Or they might expect one another to have that tendency, and act accordingly; or they might expect each other to expect each other to have that tendency and act accordingly, and act accordingly; and so on. Or – more likely – there might be a mixture of these. Their first- and higher-order expectations of a tendency to pick the salient as a last resort would be a system of concordant expectations capable of producing equilibrium at the salient equilibrium. (pp. 35–6)

Lewis thinks that coordination problems of the kind investigated by Schelling – self-contained games, without prior communication – are 'an extreme case'. However, they shed light on the 'more common case . . . of a *familiar* coordination problem without communication' (p. 36).

Lewis hypothesises that, other things being equal, individuals who face familiar coordination problems tend to follow *precedents*. But why do they have this tendency? Lewis explains the 'force of precedence' by arguing that, in any particular interaction between particular individuals, precedence is just a form of salience:

> [P]recedent is merely the source of one important kind of salience: conspicuous uniqueness of an equilibrium because we reached it last time. We may tend to repeat the action that succeeded before if we have no strong reason to do otherwise. Whether or not any of us really has that tendency, we may somewhat expect each other to have it, or expect each other to expect each other to have it, and so on . . . (pp. 36–7).

Thus, although Schelling's self-contained games, taken at face value, represent highly unusual situations, the kind of reasoning by which (on Lewis' account) people solve them is the same as that by which people

[23] This explanation is discussed in more detail by Mehta, Starmer and Sugden (1994), who argue that it is not the explanation offered by Schelling himself.

are led to reproduce coordination equilibria, once those equilibria have become established as regularities in a population.

It is not entirely clear how Lewis interprets this 'tendency' to follow precedent. On one reading, it is a non-rational psychological propensity which a person just happens to have, which governs her decisions if and when rationality gives no guidance at all. On another reading, precedent allows the individual to make inductive inferences in which she has *some* confidence, but which are overridden whenever deductive analysis points clearly in a different direction. If we accept Hume's (1740/1978, pp. 86–106; quotation from p. 103) analysis of induction – that inductive inferences are ultimately grounded on habitual associations of ideas, on 'custom acting upon the imagination' – there is not a great distance between these two interpretations. However, we suggest the second interpretation coheres better with Lewis' analysis as a whole – for two reasons. First, Lewis' concept of 'reason to believe' encompasses reasons that are grounded in inductive inferences. It would not be in the spirit of the analysis to make a sharp distinction between 'rationality' and inductive inference. Second, the idea that precedent matters only when reason is completely silent is liable to generate paradoxes.[24]

Lewis then points to the difficulties created by a feature of real-world coordination problems that most later game theorists have overlooked: that no two interactions are exactly alike.[25] Any two real-world interactions will differ in matters of detail, quite apart from the inescapable fact that 'previous' and 'current' interactions occur at different points in time.[26] Thus, the idea of 'repeating what was done in previous instances of the game' is not well-defined. Precedent has to depend on analogy: to follow precedent in the present instance is to behave in a way that is *analogous with* behaviour in past instances. In any particular case, there are (as Lewis points out) 'always innumerable possible analogies' (pp. 36–8). The fact that each interaction occurs at a specific point in time gives rise to a further difficulty. The information content of each individual's experience of a game can be represented by a *series* of observations of strategy choices

[24] Suppose precedent matters only when reason is silent. If, in a coordination problem, one player i has no reasons for choosing one strategy rather than another, he tends to follow precedent. But if another player j has reason to believe that this is in fact true for i, she *does* have a reason to follow precedent, namely, as her best reply to what she has reason to expect i to do. But then if i has reason to believe that all this is true for j, he has a reason to follow precedent, contrary to the original supposition. This paradox is discussed by Gilbert (1989, p. 74).

[25] The small minority of theorists who have considered this problem includes Sugden (1986, 1998), Goyal and Janssen (1996), and Schlicht (2000).

[26] It might seem that this feature of the real world is incompatible with a formal model of recurrent play of a *specific* game. But, in defining conventions in terms of recurrent situations, Lewis allows some fuzziness: it is not required that every instance of a given recurrent situation is identical with every other (compare fn. 7).

(and not simply by an unordered set of such observations). There are always innumerable possible ways of extending a finite series of entities. The upshot of these difficulties is that, given any finite series of 'similar' interactions observed by any individual, there is an indefinite number of patterns, each of which fits players' behaviour as so far observed, and which in principle could be interpreted as 'the precedent'. This, of course, is a familiar problem in the analysis of induction.

In general terms, the problem is this. Inductive inference is possible only because a very small subset of the set of possible patterns is privileged: only patterns from that privileged set are treated as *projectible* – as 'genuine' regularities that can be projected into a wider domain. This is the problem displayed in Nelson Goodman's (1954) famous 'grue problem'[27]; that some explanatory concepts (and hence some patterns) have to be privileged is essential to Goodman's proposed solution. This problem cannot be avoided by using Bayesian analysis. If, in a Bayesian framework, we postulate an infinite number of possible regularities, then the principle of assigning uniform priors – the principle of insufficient reason – is ill-defined. Alternatively, if we postulate an astronomically large but finite number of possible regularities, and then assume Bayesian updating of uniform priors, we run into another difficulty: even if there is a 'true' regularity with which every observation is consistent, an astronomically large number of observations will have to be made before the subjective probability of that regularity gets anywhere close to one. For Bayesian learning to work on a practical time scale, the true regularity must be assigned a higher prior probability than is warranted by the principle of insufficient reason. In other words: the true regularity must have a privileged status *prior to* the learning process.

The following passage is Lewis' response to these difficulties:

> Were it not that we happen uniformly to notice some analogies and ignore others – those we call 'natural' and 'artificial' respectively – precedents would always be completely ambiguous and worthless. . . . Fortunately, most of the analogies are artificial. We ignore them . . . And fortunately we have learned that all of us will mostly notice the same analogies. That is why precedents can be unambiguous in practice, and often are. (pp. 37–8)

Thus, for Lewis, common standards of inductive inference are ultimately grounded in common conceptions of the 'naturalness' of analogies. Or,

[27] The grue problem is this. Up to today, all the emeralds we have seen have been green. An apparently natural inductive inference is that any emeralds we see tomorrow will also be green. But consider the concept 'grue', defined so that an object is grue if and only if it is green on any day up to today and blue on any day from tomorrow. Up to today, all the emeralds we have seen have been grue as well as green. So why are we not entitled to make the inductive inference that any emeralds we see tomorrow will be grue, and hence also blue? Why is 'green' a projectible concept, and 'grue' not?

as we would prefer to say, they are grounded in common notions of projectibility.

We cannot pretend that Lewis presents a complete analysis of the role of analogy in inductive inference. To attempt such an analysis ourselves would take us far beyond the task we have set ourselves in this paper, of reconstructing Lewis' game theory. The important point is that Lewis' theory, in contrast to most modern game theory, analyses the *process of reasoning* by which each player in a game decides to play her part of an equilibrium profile of strategies. In doing so, it reveals the essential role of inductive reasoning in the reproduction of tacit conventions.

We have now arrived at a Lewisian answer to the question: where does all the iterated reason to believe come from? And, in reaching this answer, we have explained the role that salience plays in Lewis' analysis of convention. More precisely, the question we have answered is this: Given that, up to the present, behaviour in a population has exhibited some regularity which corresponds with a particular coordination equilibrium, how does it come about that there is iterated reason to believe that this regularity will persist? The answer is that in relation to A (the regularity that has been exhibited up to the present) and x (its continuation in the immediate future), the principles of reasoning endorsed by each individual satisfy conditions D1 to D5. For this to be the case, everyone must share certain common standards of inductive inference, according to which A is a privileged or projectible regularity, and each person must have reason to believe that each other person shares his own standards about what can be inferred inductively from A. On Lewis' account, projectibility (or 'natural analogy') is a form of salience.

As noted above, Lewis' definition of convention requires not only that there be iterated reason to believe that a coordination equilibrium will persist, but also that everyone in fact expects it to persist (that is, that $b_i(x)$ is true for each person i in the population), and that it does in fact persist (that is, that x is true). Given that $A = (A_1$ *and* $\ldots$ *and* $A_n)$ holds and that D1 to D5 are satisfied, we now consider what more is needed in order to ensure that $b_i(x)$ is true for each i. It is sufficient that each person i actually believes that A_i holds and follows through two steps of inductive reasoning, each of which (by assumption) is normatively legitimate within the standard of reasoning that he endorses. The first is the inference from 'A_i holds' to 'A holds'; the second is the inference from 'A holds' to x. These steps are legitimated by D3 and D4. Notice that what is required of i is only a very limited capacity to believe what he has reason to believe. In particular, no assumptions need to be made about i's actual beliefs about what other individuals believe, or have reason to believe.

Thus, Lewis' definition of convention does not require actual beliefs higher than the first order. All the iteration required is in the realm of reasons to believe, rather than in that of actual belief. Further, it does not

require individuals to hold *any* actual beliefs – not even first-order ones – about other individuals' rationality. Notice that these are statements about what is *required* for the analysis. The analysis can, of course, allow that actual beliefs may sometimes exceed these minimum requirements and include either attributions of rationality or higher-order beliefs about continuation of the regularity. However, for his part, Lewis speculates that actual beliefs of real people rarely extend beyond the fourth order – that is, in this context, beyond ones which can be represented by formulae such as $b_i[b_j[b_k[b_l[x]]]]$ (p. 32).

One question remains within Lewis' account of the reproduction of conventions: Given that $b_i(x)$ is true for each i, what more is needed to make x true? Recall that x is the continuation of a particular coordination equilibrium (driving on the right). If i believes that almost everyone else will continue to play her part in this equilibrium, the only extra condition needed to ensure that i continues to play *his* part in the equilibrium is that he is rational in the sense that his actions are governed by his desires and beliefs. And if each individual i does continue to play his part, their actions combine to make x true.

7. WHY NOT ALL COORDINATION EQUILIBRIA CAN BE TACIT CONVENTIONS

Lewis' analysis of the role of salience in the reproduction of conventions has a very important implication: not all coordination equilibria are capable of being tacit conventions. More precisely: consider any recurrent coordination problem in any population P. Consider any regularity in behaviour R such that, if everyone in P conformed to R, a coordination equilibrium would be reached in every instance of that problem. Suppose that, for some period up to the present, everyone in P has behaved consistently with R. Will this regularity persist? According to Lewis' theory, we have reason to expect R to persist as a tacit convention, only if, for members of P, R is a projectible regularity. And, in any given population at any given time, not all regularities are projectible.

For example, consider a population $P = \{i, j\}$ where i and j are two car-drivers who occasionally meet head-on on the narrow lane which links their two homes to the nearest road; if they are to pass one another, either both must steer to the left, or both must steer to the right. Each such interaction can be modelled as a coordination problem in which the strategy sets are $S_i = \{\text{left, right}\}$ and $S_j = \{\text{left, right}\}$, and in which the strategy profiles <left, left> and <right, right> are coordination equilibria. For the purposes of this example, we assume that each player conceptualises her strategies as 'left' and 'right', just as we have done as modellers. (Of course, this step itself involves an assumption, and one which is far from trivial, about shared conceptions of

salience.) Suppose that, for six interactions in succession, each individual chooses a strategy independently and at random. By pure chance, they coordinate successfully in all six interactions. Now suppose that each player, recognising this remarkable success, tries to follow the precedent set in those six interactions.

One way of representing this precedent is as a sequence of six letters, 'l' denoting 'coordination achieved by passing on the left' and 'r' denoting 'coordination achieved by passing on the right'. Given this representation, to try to follow the precedent is to look for a pre-eminently conspicuous continuation of the sequence. For a few of the 64 possible six-letter sequences, most people will probably agree that one particular continuation is uniquely salient – for example, that <r, r, r, r, r, r> is followed by r, or that <l, r, l, r, l, r> is followed by l. But what about, say, <r, l, r, r, r, l >? There is no shortage of infinite sequences which begin with this particular sequence of six letters, but none of them seems uniquely salient; in some of them the seventh letter is l, in others r. For every one of these infinite sequences, the following is true: it is a regularity; it has been followed by both members of the population up to the present, and has consistently led to coordination; and if followed in future interactions, it would continue to produce coordination. But the fact that all this is true of some particular regularity R cannot be sufficient to give i and j a compelling reason to follow it – since it is equally true of other regularities which have different continuations.

The implication, then, is that only *projectible* regularities are capable of generating the kinds of expectations which, in Lewis' analysis, allow coordination equilibria to reproduce themselves as tacit conventions. A tacit convention is (amongst other things) a regularity in behaviour which generates a pattern in each person's experience which that person treats as projectible and such that everyone's projections coincide. The set of feasible tacit conventions is constrained to be one of those patterns which, were it to be realised in every person's experience of the game, each person would treat as projectible in the same way. As a matter of logical necessity, it cannot be the case that, at any given time, *all* patterns are projectible in this sense.

8. HOW DO CONVENTIONS BEGIN?

Skyrms' second question, 'Where is the salient equilibrium?', is presented as a question about the origins of conventions (p. 84). Although we have explained the role of salience in Lewis' theory, we have so far said almost nothing about how, according to Lewis, conventions begin. There is a simple reason for this: Lewis says very little about the origin of conventions.

Skyrms implies that Lewis invokes salience to explain why one convention rather than another comes into existence – that is, to solve a problem of equilibrium selection *between alternative conventions*, viewed from some prior starting point. But in fact Lewis is not much concerned with this problem. He invokes salience as an equilibrium selection mechanism, but (apart from a few incidental remarks) only in explaining *how conventions reproduce themselves*. As we have explained, Lewis treats each instance of a recurrent coordination problem as posing its own problem of equilibrium selection; this problem is solved by the salience *of precedent*. It is only because a regularity of behaviour is already in existence that there is a precedent to follow.

We have already referred to the brief passage in which Lewis notes that, in self-contained coordination problems, coordination equilibria can be achieved by exchanges of promises, by exchanges of declarations of intent, and by the kind of salience revealed in Schelling's experiments. In another equally brief passage, Lewis suggests that self-contained signalling problems can sometimes be resolved by creative and unilateral invention of signals. His example is of someone who discovers a patch of quicksand, and, with the intention of warning others, 'puts a scarecrow up to its chest in the quicksand, hoping that whoever sees it will catch on'. The idea, presumably, is that a stylised representation of a human figure submerged up to its chest will, by a natural association of ideas, prompt thoughts about real human beings being similarly submerged. If this attempt at signalling works, it does so by means of salience (pp. 158–9). This is just about all that Lewis has to say about how coordination equilibria might be reached, other than by precedent. It is perhaps reasonable to infer that Lewis thinks that the mechanisms described in these passages *could* work as the first stages in the emergence of conventions, but he never actually says this outright. He certainly does not make any general claim to the effect that real conventions *in fact* have their origins in these mechanisms.

Lewis' game theory, like almost all game theory at the time he was writing, is primarily concerned with equilibrium. From the standpoint of modern evolutionary game theory, such a theory is incomplete: it needs to be supplemented by an analysis of the dynamic processes by which equilibria come about. Even so, Lewis' analysis of how tacit conventions reproduce themselves has implications for the question of which tacit conventions emerge. If his analysis is correct, the answer to this question cannot be (as Skyrms claims it can be) entirely a matter of chance, independent of salience.

In Skyrms' theory, tacit conventions originate as asymmetries that are generated by chance in random processes and are then amplified by differential replication. Theories of this kind, which are becoming

widely accepted among evolutionary game theorists,[28] are clearly useful in capturing some aspects of evolutionary processes. But, in any evolutionary theory, what is capable of being replicated must depend on facts about the mechanism of replication. Skyrms, like most other social theorists who use evolutionary game theory, interprets the process of replication as working through mental processes such as imitation and reinforcement learning.[29] Any process that involves the imitation or learning of successful patterns of behaviour requires a prior mechanism for *recognising* patterns, whether successful or not: patterns that are not recognised cannot be imitated or learned. If, at any given time, not all conceivable regularities are projectible, not all conceivable strategies are capable of being imitated or learned. In random processes of the kind considered by Skyrms, the vast majority of the asymmetries in behaviour that these processes can generate are not projectible. Only the projectible asymmetries have the potential to evolve into tacit conventions.

In any satisfactory model of the evolution of tacit conventions, the evolutionary path starting from any point in time must be constrained by whatever shared conceptions of salience are available at that time. This is not to say that conceptions of salience must be treated as constant. To the contrary, they evolve over time too. As people discover new regularities in their experiences, they may come to recognise new analogies and new precedents; patterns which previously were not considered as projectible may become so, and conversely. Thus, conceptions of salience tend to track evolving conventions. (For example, consider how languages evolve. In modern English, plurals are almost always formed by adding '-s' to singular nouns. For modern English speakers, this is a highly salient precedent, which ensures that new nouns almost invariably come to have an '-s' plural. But the current salience of '-s' plurals is the product of an evolutionary history. At one time, English had many more plurals in '-en' than it does now; we can assume that as these older forms were gradually displaced, the salience of '-s' increased.[30]) It seems that we need a theory of the co-evolution of conventions and of conceptions of salience.[31]

In such a theory, the reproduction of conventions may be explained in terms of conceptions of salience, while conceptions of salience may be

[28] Young (1998) presents a formal theory of such processes.

[29] Skyrms (1996) does not say much about how the processes of replication in his models are to be understood. But when he first introduces the idea of evolutionary modelling, he suggests the interpretation that strategies are more likely to be imitated, the more successful they are (pp. ix–xi).

[30] Strang (1970) describes and analyses many such evolutionary processes in the history of the English language. In building theories of social evolution, economists have tended to draw on analogies in biology, but for understanding the role of salience, analogies with linguistics might also prove fruitful.

[31] Sugden (1986) gives a rough sketch of such a theory.

explained in terms of current conventions. Is this circular? Logically, no, because these explanations respect the arrow of time: later phenomena are always being explained in terms of earlier phenomena. Further, such explanations can allow the slightly more complex to be explained in terms of the slightly less, so gradual accumulations of complexity can be explained. The structure of such explanations is similar to that of many explanations in biology: the Darwinian theory of natural selection is enormously powerful in explaining regularities among living things, despite the fact that the ultimate origin of life is poorly understood.

Even so, a critic might object that explanations of this kind do not go deep enough, that they leave open fundamental questions about the origins of conventions that ought to be answered. In the case of language, Skyrms makes an objection of this kind when he says that Lewis' theory does not answer Quine's scepticism. Skyrms, it seems, thinks that a satisfactory Humean reply to Quine ought to show that conventions of language can emerge among individuals who initially have *no language at all*. According to Skyrms, it is question-begging in this context to assume any kind of mutual understanding among individuals, unless it can positively be shown that that mutual understanding is possible without language. Clearly, Lewis' theory does not meet this criterion for a satisfactory reply to the sceptics. Lewis is not trying to explain the origins of language, but only to provide an analysis of convention which, as he puts it, 'permits language to be conventional' (p. 2). He claims to show that signalling conventions can emerge through the formation of common expectations, and that this process can be entirely tacit: it need not involve any use of language. He neither asserts nor denies that the process is possible among beings which have no language.

Our purpose in this paper has been to reconstruct and extend Lewis' game theory, in relation to common knowledge and convention. We do not address the deep question of whether any Humean theory of language can satisfy Skyrms' criterion. However, Lewis' analysis does tell us something about what such a theory would have to be like, if it were to satisfy that criterion. Since it could not dispense with assumptions about common standards of inductive inference, it would have to show that, in the absence of language, at least some inductive standards could be common – and could reasonably be believed to be common – among the members of a human population.[32]

[32] One possible starting point would be to show that human beings are born with innate tendencies to privilege certain patterns when making inductive inferences. In fact, there is evidence that very young babies are predisposed to recognise some patterns and not others – for example, that they are predisposed to recognise vertical symmetry, classifying two objects that are alike except for left-right reflection as 'the same' (Mehler and Dupoux, 1994). It is easy to see how this principle of classification helps to adapt human beings to a world in which many significant natural objects have vertical symmetry.

9. CONCLUSION

Lewis' *Convention* is generally recognised to be a major contribution to philosophy and to game theory. But the more closely we have read the text, the more we have discovered. Behind the apparently informal exposition there is an astonishingly comprehensive, consistent and original theory of practical reason in the context of recurrent games. Lewis' approach offers an alternative both to the modelling strategy of classical game theory, in which self-contained games are played by hyper-rational agents, and to that of evolutionary game theory, in which players' behaviour is the product of blind processes of selection. We hope that, by offering a formal reconstruction and extension of Lewis' analysis, we have shown that it is not vulnerable to the criticism that it simply assumes the existence of high-order iterated beliefs, nor to the criticism that its assumptions about salience are redundant. To the contrary: to understand Lewis' approach is to recognise the limitations both of standard treatments of common knowledge and of current attempts to construct a social evolutionary game theory in which salience plays no part.

APPENDIX 1: PROOFS OF THEOREMS

Proof of Lewis' Theorem

Consider any state of affairs A, any proposition x, and any population P. Suppose that A holds and that A is a reflexive common indicator in P that x. Then:

1.	For all i in P: R_i(A holds)	(from C1)
2.	For all i, j in P: A ind_i R_j(A holds)	(from C2)
3.	For all i in P: A ind_i x	(from C3)
4.	For all i in P: R_i(x)	(from 1 and 3, using A1)
5.	For all i, j in P: R_i(A ind_j x)	(from 3, using C4)
6.	For all i, j in P: A ind_i R_j(x)	(from 2 and 5, using A6)
7.	For all i, j in P: $R_i[R_j(x)]$	(from 1 and 6, using A1)
8.	For all i, j, k in P: R_i[A ind_j R_k(x)]	(from 6, using C4)
9.	For all i, j, k in P: A ind_i $R_j[R_k(x)]$	(from 2 and 8, using A6)
10.	For all i, j, k in P: $R_i[R_j(R_k[x])]$	(from 1 and 9, using A1)
11.	For all i, j, k in P: R_i[A ind_j $R_k(R_l[x])$]	(from 9, using C4)

And so on. The role played by 'x' in lines 3–5 is played by 'reason to believe x' in lines 6–8, by 'reason to believe reason to believe x' in lines 9–11, and so on. Lines 4, 7, 10, . . . establish the theorem.

Proof of Distribution Theorem

Consider any proposition x, any population $P = \{1, \ldots, n\}$, and any state of affairs $A = (A_1$ *and* $\ldots$ *and* $A_n)$, such that A is a distributed indicator in P that x. Then:

1.	For all i in P: (A holds) entails (A_i holds)	(from definition of A)
2.	For all i in P: (A holds) $\Rightarrow$ R_i(A_i holds)	(from 1, using D1)
3.	For all i, j in P: A_j ind_i R_j(A_j holds)	(from D2)
4.	For all i, j in P: A_j ind_i (A holds)	(from D3)
5.	For all i in P: A ind_i x	(from D4)
6.	For all i in P: (A holds) $\Rightarrow$ R_i(A holds)	(from 2 and 4, using A1)
7.	For all i, j in P: A ind_i (A_j holds)	(from 1, using A2)
8.	For all i, j in P: R_i[A_j ind_j (A holds)]	(from 4, using D5)
9.	For all i, j in P: A_j ind_i R_j(A holds)	(from 3 and 8, using A6)
10.	For all i, j in P: A ind_i R_j(A holds)	(from 7 and 9, using A4)

Lines 5, 6 and 10 respectively establish that C3, C1 and C2 hold for A. D5 entails that C4 holds for A. Thus, A is a reflexive common indicator in P that x.

Proof of Iterated Belief Theorem

Consider any state of affairs A, any proposition x, and any population P, such that (i) A is a reflexive common indicator in P that x and (ii) A is a reflexive common indicator in P that each person in P reasons faultlessly. Suppose that A holds and that each person in P reasons faultlessly. Then:

1.	For all i in P: R_i(A holds)	(from C1)
2.	For all i, j in P: A ind_i R_j(A holds)	(from C2)
3.	For all i in P: i reasons faultlessly	(supposition of theorem)
4.	For all i, j in P: A ind_i (j reasons faultlessly)	(from C3)
5.	For all i in P: A ind_i x	(from C3)
6.	For all i in P: R_i(x)	(from 1 and 5, using A1)
7.	For all i in P: b_i(x)	(from 3, 6 and definition of 'faultless reasoning')
8.	For all i, j in P: R_i(A ind_j x)	(from 5, using C4)
9.	For all i, j in P: A ind_i R_j(x)	(from 2 and 8, using A6)
10.	For all i, j in P: A ind_i [R_j(x) $\wedge$ (j reasons faultlessly)]	(from 4 and 9, using A3)
11.	For all j in P: [R_j(x) $\wedge$ (j reasons faultlessly)] entails b_j(x)	(from definition of 'faultless reasoning')
12.	For all i, j in P: A ind_i b_j(x)	(from 10 and 11, using A5)
13.	For all i, j in P: R_i[b_j(x)]	(from 1 and 12, using A1)

14.	For all i, j in P: $b_i[b_j(x)]$	(from 3, 13 and definition of 'faultless reasoning')
15.	For all i, j, k in P: $R_i[A\ ind_j\ b_k(x)]$	(from 12, using C4)
16.	For all i, j, k in P: $A\ ind_i\ R_j[b_k(x)]$	(from 2 and 15, using A6)

And so on. The role played by 'x' in lines 5–9 is played by 'belief that x' in lines 12–16 and so on. Lines 7, 14, . . . establish the theorem.

APPENDIX 2: AUMANN'S MODEL OF COMMON KNOWLEDGE, AND VANDERSCHRAAF'S RECONSTRUCTION OF LEWIS' ANALYSIS

In the model of common knowledge presented by Aumann (1987), there is a set P of individuals, each of whom is rational in the Bayesian sense. There is a finite set Ω of *states* (of the world), understood as complete descriptions of how the world might be. One and only one of these states *obtains*, i.e. describes how the world actually is, but individuals do not necessarily know which state this is. Sets of states are *events*; an event obtains if any of its elements obtains. Each individual assigns a *prior* subjective probability to each state. Subjective probabilities conditional on events are formed by Bayesian revision of priors. For each individual i there is an *information partition* I_i of Ω. This represents what i knows in each state. That is: if E is an element of I_i, and if some state ω in E obtains, then i knows that E obtains (and hence that each superset of E obtains). The profile of individuals' information partitions is the *information structure*.

Notice that this model makes a sharp distinction between knowledge and belief. Knowledge is represented by information partitions, belief by subjective probabilities. What an individual knows is true, by virtue of the properties of the model, but beliefs are constrained only by those requirements of internal consistency imposed by the Bayesian calculus of probability. Aumann defines common knowledge in terms of information partitions. Thus, for Aumann, 'common knowledge' is genuinely a matter of knowledge, while for Lewis it is a matter of reason to believe.

Aumann's (1987, pp. 8–12; 1999, pp. 272–3, 276–8) interpretation of his model requires that each individual knows all the properties of the model, i.e. the possible states of the world (as specified by Ω), every individual's information partition, every individual's priors, and the fact that every individual is rational in the Bayesian sense. Further, each individual i knows that each individual j knows this, and so on. Here, we are using 'know' in its everyday sense: we are discussing the interpretation of the model, not its formal structure. Within the model, 'know' has a technical meaning. (From now on, whenever we use *know* in this technical sense, we shall use italics.) Thus, Aumann's model represents a world in which everything apart from which state actually obtains is, in the informal sense of 'knowing', already common knowledge. The existence of common

knowledge in this sense is a datum; Aumann's theory is not intended to explain how it comes about.

Defending this approach, Aumann argues that these assumptions do not have any substantive content, but merely represent a modelling strategy. The idea is this: whatever substantive assumptions we (as modellers) want to make about individuals' knowledge or lack of knowledge, those assumptions can be represented in *some* model of the kind Aumann proposes. That is, we can represent them in terms of *some* specification of states of the world and information partitions, such that this specification is, in the informal sense, common knowledge. Notice, however, that this defence justifies Aumann's modelling strategy as a way of *representing* what individuals know. The strategy cannot be used to *explain* how those individuals come to know what they are represented as knowing. In contrast, Lewis' theory is intended as an analysis of the processes of reasoning by which individuals can form beliefs about the world.

Aumann defines a *knowledge* operator K_i such that, for all individuals i, for all events $A \subseteq \Omega$:

$$K_i(A) = \{\omega \in \Omega \mid (\exists E \in I_i)(\omega \in E \wedge E \subseteq A)\}$$

This is read as 'i *knows* A'. Notice that $K_i(A)$ *is an event*: it is the set of all those states ω with the property that, if ω obtains, i knows (from his information partition) that A obtains.

Because both $K_i(A)$ and A are events, *knowledge* operators can be nested to any finite depth. Thus, for example, $K_j[K_i(A)]$, read as 'j *knows* that i *knows* A', is an event. At first sight, this reading might seem to conflate extensionality and intensionality in an illegitimate way. Suppose we (the modellers) know that the true state ω is an element of $K_j[K_i(A)]$. We are entitled to conclude that there is *some* set of states, which we may denote by A', such that j knows that A' obtains, and such that A' contains all those states, and only those states, at which i knows that A obtains. In other words, A' is extensionally equivalent to 'i knows that A obtains', and *we* know that. But does j? What licenses the transition to 'j knows that i knows that A obtains'? This transition is valid only on the assumption that j knows that A' is extensionally equivalent to 'i knows that A obtains'.

Now consider the event $K_k[K_j[K_i(A)]]$. Suppose that we (the modellers) know that the true state ω is an element of that event. Then we are entitled to conclude that there are sets of states A' and A'', such that k knows that A'' obtains, A'' is extensionally equivalent to 'j knows that A' obtains', and A' is extensionally equivalent to 'i knows that A obtains'. If we are to interpret $K_k[K_j[K_i(A)]]$ as 'k knows that j knows that i knows that A obtains', we need to assume that k knows that j knows that A' is extensionally equivalent to 'i knows that A obtains'. And so on. Thus, if *knowledge* operators can be nested to any finite depth, we need to assume that certain properties of extensional equivalence are (in the informal sense) common

knowledge. This assumption has to be made outside the formal model, because the *knowledge* operator applies only to events, and propositions about the extensional equivalence of different descriptions of events are not themselves events. It is a consequence of Aumann's more general background assumption that the properties of the model are common knowledge.

Aumann defines an event A to be common *knowledge* in a state ω if ω is an element of all events of the (finite) form $K_i[K_j[\ldots K_k(A)\ldots]]$. Thus, if A is common *knowledge* in ω and if ω obtains, then i *knows* that j *knows* that ... that k *knows* A. On this analysis, common *knowledge* of an event is a specific property of an information structure in a model in which the information structure itself is, by assumption, common knowledge.

Vanderschraaf (1998b, pp. 361–3) reconstructs Lewis' analysis of common knowledge, using a set-theoretic *knowledge* operator similar to Aumann's. Like Aumann, he treats *knowledge* formulae as events, and uses nested *knowledge* formulae to represent within his model what one person knows about what another person knows. However, he does not assume that there is an information partition for each person. Instead, he assumes that the *knowledge* operator satisfies the following four conditions:

(V1) For all persons i in P, for all events A: $K_i(A) \subseteq A$.
(V2) For all persons i in P: $\Omega \subseteq K_i(\Omega)$.
(V3) For all persons i in P, for all sets of events $\mathcal{A}$: $K_i[\cap_{A\in\mathcal{A}}] = \cap_{A\in\mathcal{A}} K_i(A)$.
(V4) For all persons i in P, for all events A: $K_i(A) \subseteq K_i[K_i(A)]$.

These conditions are satisfied by Aumann's *knowledge* operator, but they can also be satisfied by *knowledge* operators that cannot be represented by information partitions.

Vanderschraaf represents Lewis' 'i has reason to believe that A holds' as $K_i(A)$, i.e. as 'i *knows* A'. This conflates knowledge and reason to believe (so, for example, making it impossible to state the possibility that a person has reason to believe something false) and eliminates what we take to be an important feature of Lewis' analysis, namely its concern with *reasoning*. Vanderschraaf represents Lewis' 'A indicates to i that A′ holds' as $K_i(A) \subseteq K_i(A')$, i.e. as 'if i *knows* A, then i *knows* A''. This has the effect of translating Lewis' *if ... thereby ...* formula as a material implication between events, rather than (as in our reconstruction) as an inference rule in a logic of reasoning.

Reconstructing Lewis' definition, Vanderschraaf defines an event A to be common knowledge in a state ω if there exists some event A* such that ω is in A* and the following conditions are satisfied:

(V5) For all persons i in P: $\omega \in K_i(A^*)$.
(V6) For all persons i in P: $K_i(A^*) \subseteq K_i[\cap_{j\in P} K_j(A^*)]$.

(V7) For all persons i in P: $K_i(A^*) \subseteq K_i(A)$.
(V8) For all persons i, j in P, for all states of affairs A: $[K_i(A^*) \subseteq K_i(A) \wedge K_i(A^*) \subseteq K_i(K_j[A^*])] \Rightarrow K_i(A^*) \subseteq K_i(K_j[A])$.

Here, A* corresponds with what, in our reconstruction of Lewis' analysis, we represent as a reflexive common indicator in P that A holds. Vanderschraaf describes V8 as requiring that individuals are 'symmetric reasoners', but, strictly speaking, it is a property of events rather than of reasoning. Using these conditions, Vanderschraaf proves that if A is common knowledge in ω according to this definition, then it is also common *knowledge* in ω according to Aumann's definition.

The interpretation of Vanderschraaf's model requires implicit assumptions about common knowledge, for essentially the same reasons that the interpretation of Aumann's does. (Notice that our discussion of the problem of extensional equivalence in Aumann's model makes no reference to information partitions; thus, it applies with equal force to Vanderschraaf's model.) Thus, Vanderschraaf's reconstruction of what we have called Lewis' Theorem cannot be interpreted as an answer to the question of how iterated belief or iterated reason to believe comes about. If we want an answer to that question, we cannot use a modelling strategy which presupposes that properties of the model are common knowledge.

REFERENCES

Aumann, Robert J. 1976. Agreeing to disagree. *Annals of Statistics* 4:1236–9
Aumann, Robert J. 1987. Correlated equilibrium as an expression of bayesian rationality. *Econometrica* 55:1–18
Aumann, Robert J. 1999. Interactive epistemology I: Knowledge. *International Journal of Game Theory 28* :263–300
Bacharach, Michael. 1992. The acquisition of common knowledge. In *Knowledge, belief and strategic interaction*, Cristina Bicchieri and M. L. Dalla Chiara (eds.), 285–315. Cambridge University Press
Binmore, Ken. 1998. *Just playing*. MIT Press
Gilbert, Margaret. 1989. Rationality and salience. *Philosophical Studies* 57:61–77
Goodman, Nelson. 1954. *Fact, fiction, and forecast*. Harvard University Press
Goyal, Sanjeev, and Maarten Janssen. 1996. Can we rationally learn to coordinate? *Theory and Decision* 40:29–49
Hume, David. [1740] 1978. *A treatise of human nature*. Oxford University Press
Lewis, David. 1969. *Convention: A philosophical study*. Harvard University Press
Maynard Smith, John, and G. R. Price. 1973. The logic of animal conflicts. *Nature* 246:15–8
Mehler, Jacques and Emmanuel Dupoux. 1994. *What infants know: The new cognitive science of early development*. Blackwell
Mehta, Judith, Chris Starmer, and Robert Sugden. 1994. The nature of salience: an experimental investigation of pure coordination games. *American Economic Review* 84:658–73
Nash, John. 1950. Non-cooperative games. Ph.D. diss, Princeton University
Nozick, Robert. 1963. The normative theory of individual choice. Ph.D. diss, Princeton University

Nozick, Robert. 2001. *Invariances: The structure of the objective world*. Harvard University Press
Quine, W. V. O. 1936. Truth by convention. In *Philosophical essays for A. N. Whitehead*, O. H. Lee (ed.). Longmans
Ritzberger, Klaus, and Jörgen Weibull. 1995. Evolutionary selection in normal-form games. *Econometrica* 63:1371–99
Russell, Bertrand. 1921. *The analysis of mind*. Allen and Unwin
Samuelson, Larry. 1997. *Evolutionary games and equilibrium selection*. MIT Press
Savage, Leonard. 1954. *The foundations of statistics*. Wiley
Schelling, Thomas. 1960. *The strategy of conflict*. Harvard University Press
Schiffer, Stephen. 1972. *Meaning*. Oxford University Press
Schlicht, Ekkehart. 2000. Aestheticism in the theory of custom. *Journal des Economistes et des Etudes Humaines 10*:3–51
Skyrms, Brian. 1990. *The dynamics of rational deliberation*. Harvard University Press
Skyrms, Brian. 1996. *Evolution of the social contract*. Cambridge University Press
Skyrms, Brian. 1999. Evolution of inference. In *Dynamics in human and primate societies*, T. Kohler and G. Gumerman (eds.), 77–88. Oxford University Press
Strang, Barbara. 1970. *A history of English*. Methuen
Sugden, Robert. 1986. *The economics of rights, cooperation and welfare*. Blackwell
Sugden, Robert. 1998. The role of inductive reasoning in the evolution of conventions. *Law and Philosophy* 17:377–410
Vanderschraaf, Peter. 1995. Convention as correlated equilibrium. *Erkenntnis* 42:65–87
Vanderschraaf, Peter. 1998a. The informal game theory in Hume's account of convention. *Economics and Philosophy* 14:215–47
Vanderschraaf, Peter. 1998b. Knowledge, equilibrium and convention. *Erkenntnis* 49:337–69
Weibull, Jörgen. 1995. *Evolutionary game theory*. Harvard University Press
White, Morton. 1950. The analytic and the synthetic: an untenable dualism. In *John Dewey: Philosopher of science and freedom*, Sidney Hook (ed.). Dial
Young, H. Peyton. 1993. The evolution of conventions. *Econometrica* 61:57–84
Young, H. Peyton. 1998. *Individual strategy and social structure*. Princeton University Press