How inductive inferences can be grounded on salience alone:

some reflections on the emergence of conventions

Robert Sugden

School of Economics, University of East Anglia

Conventions fall into the domains of both philosophy and economics. In the study of convention, the interplay of the two disciplines has proved to be productive. As its practitioners often acknowledge, many of the key ideas in the modern theory of convention were first proposed by David Hume in his *Treatise of Human Nature* (1739–40/ 1978). Although Hume's *Treatise* is now usually classified as one of the great works of philosophy, it can also be read as an investigation in cognitive psychology which (among many other achievements) lays the foundations for an empirically-grounded form of decision theory.[1] The founding text for the modern theory of convention is David Lewis's *Convention: A Philosophical Study* (1969). As the subtitle suggests, and as the introduction to the book makes clear (pp. 1–4), Lewis presents his ideas as a contribution to philosophy: he offers an 'analysis' of convention in the sense of analytical philosophy, 'along the lines of Hume's [theory]'. But the 'source' of these ideas is 'the theory of games of pure coordination', as developed in economics by Thomas Schelling (1960). In the course of his study of convention, Lewis makes major contributions to game theory. As well providing what is generally credited as the first rigorous analysis of common knowledge, *Convention* initiates the analysis of games that are played recurrently within a population – a key ingredient of evolutionary game theory, which is now widely used in economics.[2]

These introductory comments are written with self-serving intent, to try to counter the scepticism with which readers will react to an economist's claim to have something new to say about one of the classic problems of philosophy. In this paper, I argue that the game-theoretic analysis of convention throws light on the problem of understanding the nature of inductive inference and the sense (if any) in which such inferences can be treated as valid. This *problem of induction* has usually been viewed in the perspective of the philosophy of natural science. In that perspective, it is natural to think of inductive inference as an attempt

to gain knowledge about a world whose properties are independent of how anyone thinks about them. Thus, it is tempting to think, standards of valid inductive inference ought somehow to reflect the structure of the world we are trying to gain knowledge about. (That we cannot claim to know what that structure might be without first making inductive inferences is, of course, part of the problem of induction.) But when we try to make inductive inferences about the conventions to which people like ourselves conform, there is a sense in which we are constructing the reality about which we are seeking knowledge. What difference does that make? That question initiated the train of thought that I report in this paper.

1. Lewis's theory of convention

My starting point is Lewis's theory of convention, the key properties of which I now summarise. In the interests of clarity, I leave out many features of the theory which, although important for Lewis's project as a whole, are orthogonal to the issues I am addressing. I begin with some definitions. These do not correspond exactly with Lewis's own definitions, but they are close enough for my purposes.

The analysis is of behaviour within a *population* of n individuals. Repeatedly, pairs of individuals drawn from this population engage in some game-like *interaction*. The assumption that interactions involve exactly two individuals is mine, and is made only to keep things simple; the analysis can easily be extended to interactions of three or more individuals. In some of Lewis's 'sample conventions', such as one which starts from the idea that 'you and I must meet every week', the same two individuals interact repeatedly; these correspond with $n = 2$. In other examples, such as a convention among the inhabitants of a town for dealing with telephone calls that are suddenly cut off, n is much larger (pp. 42–51). It is convenient to reserve Lewis's term *recurrent* for this latter form of population-wide repetition of an interaction, in which different instances of interaction may involve different individuals. I will focus on recurrent interactions.

In an interaction between two agents, each of them chooses from a set of mutually exclusive and exhaustive *actions*. A pair (a_1, a_2) of actions, one action for each agent, is a *coordination equilibrium* if two conditions are satisfied. First, each agent is (strictly) motivated to perform his component of the pair, conditional on his believing that the other agent will perform hers. Second, there is at least one other pair (a_1', a_2') of actions (with a_1'

$\neq a_1$ and $a_2' \neq a_2$) for which the first condition is true.[3] An interaction is a *coordination problem* if it contains at least two coordination equilibria, and if, as Lewis phrases it, 'coincidence of interest predominates' or, equivalently, 'there is no considerable conflict of interest' (pp. 16, 24). The essential idea here is that the outcomes of the interaction can be partitioned into those that occur when the agents reach a coordination equilibrium and those that result when they don't. Each agent has a strong preference for each of the former outcomes to each of the latter, while being close to indifferent between different elements within each of the two sets separately.

Suppose that some coordination problem is faced recurrently in some population *P*. Let *R* be some population-wide regularity in behaviour by which agents reliably pick a particular one of the coordination equilibria. Such a regularity, if it in fact occurs, is a *convention*. Considered merely as a possible state of affairs, it is a *putative convention*. (Thus, a recurrent coordination problem entails the existence of two or more putative conventions. At most one of these is realised as an actual convention.)

Take Lewis's example of the phone calls. He tells us that, in his home town of Oberlin, Ohio, there was a time when local calls were cut off without warning after three minutes. Here *P* is the population of phone users in Oberlin, and an interaction occurs between two phone users when a call between them is cut off. One putative convention is that the original caller calls back. Another is that the person called calls back. Lewis tells us that the first of these was in fact the convention.

Here is one of my favourite examples. Consider the behaviour of the drivers of two vehicles approaching one another on collision courses at a crossroads. Here *P* is the population of drivers in some geographical area. One putative convention is that the driver who is on course to reach the intersection first maintains speed, while the other gives way. Another is that the driver on the less important road gives way to the driver on the more important road. A third is that the driver who is approaching from the left gives way to the driver approaching from the right. Each of these practices has been a convention in some time and place; if one starts driving in an unfamiliar country, it is not as easy as it might seem to work out whether a convention is in operation, and if so, what it is (Sugden, 1998; 2004, pp. 36–57).

Lewis's main concern is with how, once they are in existence, conventions reproduce themselves. In his analysis, the regularity of behaviour that constitutes a convention is

supported by *concordant mutual expectations* (pp. 24–36). The essential feature of such expectations is that each member of the population expects each other member to behave in accordance with the relevant regularity. It follows immediately from the definition of a convention that if every individual has this expectation, each is motivated to behave in a way that confirms the expectations of the others. So the problem is to explain how concordant mutual expectations are created and sustained. Lewis recognises that, in many real-world cases (particularly when the relevant population is small) concordant expectations are formed by explicit agreement. But he is more interested in other cases:

> But agreement (literally understood) is not the only source of concordant expectations to help us solve our coordination problems. We do without agreement by choice if we find ourselves already satisfied with the content and strength of our mutual expectations. We do without it by necessity if we have no way to communicate, or if we can communicate only at a cost that outweighs our improved chance of coordination (p. 35).

The crossroads exemplifies the problem of being unable to coordinate expectations by explicit agreement. Typically, the two drivers do not know one another's identities and could not have agreed on a priority rule ahead of their interaction. By the time they become aware of their coordination problem, it is too late for useful communication.

Lewis begins by considering the apparently abnormal case of two individuals facing a *novel* coordination problem, unable to communicate. He notes that this case has been investigated experimentally by Schelling (1960), and glosses Schelling's findings as: 'They [i.e. Schelling's experimental subjects] try for a coordination equilibrium that is somehow *salient*: one that stands out from the rest by its uniqueness in some conspicuous respect.' In one of his many contributions to game theory, Lewis offers the following explanation of Schelling's result.[4]

> 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 each other 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 coordination at the salient equilibrium. (pp. 35–36)

At first sight, this discussion of novel coordination problems might seem to be a digression. Surely, one might think, individuals who have experience of an ongoing convention do not need to rely on salience: their can ground their expectations on *precedent*.

Lewis recognises that, in such cases, precedent is the source of agents' concordant expectations, but he wants an analysis of how agents *have reason* to base their expectations on precedent. It is here that the problem of induction begins to impinge on the theory of convention.

Lewis argues that precedent is a form of salience:

> We can explain the force of precedent just as we explained the force of salience. Indeed, precedent 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 this tendency, we may somewhat expect each other to have it, or expect each other to expect each other to have it, and so on – that is, we may each have first- and higher-order expectations that the others will do their parts of the old coordination equilibrium, unless they have reason to do otherwise. (pp. 36–37)

But, of course, no two coordination problems are exactly alike. The idea of following precedent cannot be understood as literally *repeating* a previous action. It must involve recognising some *similarity* between the present coordination problem and previous ones, and choosing a current action that is similar to an action that proved successful in those previous problems:

> So suppose not that we are given the original problem again, but rather that we are given a new coordination problem analogous somehow to the original one. Guided by whatever analogy we notice, we tend to follow precedent by trying for a coordination equilibrium in the new problem which uniquely corresponds to the one we reached before. (p. 37)

But what if we notice more than one analogy? Following precedent cannot be interpreted simply as using *the* analogy that one happens to notice; it must involve some discrimination between alternative analogies:

> In fact there are always innumerable alternative analogies. Were it not that we happen uniformly to notice some analogies and ignore others – those we call 'natural' or 'artificial', respectively – precedents would always be completely ambiguous and worthless. ... Fortunately, most of the analogies are artificial. We ignore them; we do not tend to let them guide our choice, nor do we expect each other to have any such tendency, nor do we expect each other to expect each other to, and so on. 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. 37–38)

Lewis now has the resources to explain how, when agents form reasonable beliefs and act on them, conventions reproduce themselves. Suppose that you and I are residents of

Oberlin, Ohio, three minutes into a phone call which you initiated. We are cut off. What should I do?

I notice an analogy between this event and previous cases in which calls have been cut off. Thinking about my behaviour in those previous cases, I notice a regularity: when I was the original caller, I called back, and when I was not, I waited. This behaviour has always resulted in a smooth re-establishment of communication. So I infer that other people's behaviour has exhibited the same regularity as my own. I can think of other analogies, and of other ways of describing my experience of previous phone calls, but none as natural as the one I have just described. My provisional conclusion is that, unless I can find some definite reason to do otherwise, I should stick with my previous practice and wait for you to call back.

As a first step in checking for countervailing reasons, I think about how this coordination problem appears to you. Since you too are an Oberlin resident, your experience of cut-off phone calls is likely to have been similar to mine, and your behaviour is likely to have followed the pattern that I have found in Oberlin people in general. Since different people's conceptions of naturalness in analogies tend to be in agreement, you are likely to interpret that experience in the same way that I do. I infer that you have probably reached the provisional conclusion that you should stick with your previous practice, and (since in this case you were the original caller) call back. That gives additional support to my own provisional conclusion in favour of waiting. As a further check, I might think about how you might think about how the problem appears to me, and so on. Every such check gives more support to my provisional conclusion; I find no countervailing reasons. Thus, I have reason to follow what I perceive as precedent. If everyone does this, the existing convention reproduces itself.

Lewis sums up this analysis by saying that 'coordination by precedent' amounts to 'achievement of coordination by means of shared acquaintance with a *regularity* governing the achievement of coordination in a class of past cases which bear some conspicuous analogy to one another and to our present coordination problem' (p. 41). Notice that Lewis is not trying to explain why conventions exist in the first place. His analysis is of a state of affairs in which 'conforming action produces expectation of conforming action and expectation of conforming action produces conforming action'. Or: 'And so it goes – we're here because we're here because we're here because we're here' (pp. 41–42).

For my present purposes, the significance of this argument is its analysis of inductive inference. According to Lewis, conventions reproduce themselves by the force of precedent, and precedent works through individuals' use of shared standards of inductive inference. Although Lewis uses the framework of rational-choice theory, he does not purport to demonstrate the rationality of inductive inferences. Instead, he relies on the empirical premise of a human tendency to act on a certain kind of inductive reasoning – that is, to repeat previously successful actions – when there is 'no strong reason to do otherwise'. There is no attempt to *justify* this tendency as rational. (Of course, most people's experience will tell them that, in most cases, reasoning in this way has been successful in the past. But to use this to try to justify induction would be to commit one of the most famous errors in philosophy.) In the terms used by Nelson Goodman (1954), Lewis does not try to solve the 'old problem of induction'.

Crucially, however, Lewis *does* take on Goodman's 'new' problem of induction – that of 'defining the difference between valid and invalid predictions' (Goodman, p. 65). (Since it is now over half a century since the problem was stated, I will call it Goodman's Problem.) A solution to this problem requires a general principle for distinguishing between those observed regularities that are deemed to be projectible and those that are not; but that principle is not required to be supported by a proof of the validity of the inferences it legitimates. Goodman's *grue* example illustrates the difficulty of solving even his Problem. Suppose that all emeralds examined up to time *t* have been found to be green. Do the rules of inductive inference support the conclusion that all emeralds are green? And if so, what about the fact that all emeralds examined up to time *t* have also been grue, where 'grue' is defined so that an object is grue if was found to be green up to time *t*, or if it was found to be blue after that time? The same problem occurs for Lewis's example of the phone calls. Suppose I am an Oberlin resident. Up to now (time *t*), all cut-off phone calls have been restored by the caller calling back. Why do I infer that the regularity is 'caller calls back' rather than 'up to time *t*, caller calls back; after time *t*, caller waits'? Lewis's answer is that, for an agent facing a coordination problem, a regularity in the behaviour of people who have faced similar problems in the past is projectible to the extent that the projection uses analogies *that the agent perceives as natural*.

Lewis claims that, in real-world practical reasoning, individuals tend to use this principle of inductive inference as a default. He also claims that different individuals' perceptions of naturalness tend to have significant features in common. Thus, conventions

tend to reproduce themselves. As a result, the projections that the principle legitimates are confirmed. One might expect that outcome to provide psychological reinforcement for each agent's future use of the principle. Since this feedback loop reflects specific properties of coordination problems and conventions, it is not obvious that Lewis's criterion of projectibility is a credible solution to the general problem presented by Goodman. Still, I believe that it is.

2. A pragmatic rendering of Lewis's analysis

For those who see game theory as an a priori analysis of the beliefs and actions of ideally rational agents, it may seem unsatisfactory that Lewis's explanation of the reproduction of conventions depends on an empirical assumption about a non-rational (although not *ir*rational) human propensity – the propensity to follow precedent 'if we have no strong reason to do otherwise'.[5] This, one might say, is the price that has to be paid for including a hypothesis about inductive inference in the deductive framework of rational-choice theory. And that prompts the question of whether Lewis's argument might be reformulated in terms of a more pragmatic conception of rationality.

As I have already said, Lewis was strongly influenced by Schelling's earlier analysis of coordination games. One of Schelling's most important insights, and one that Lewis uses in his analysis of precedent, is that in solving novel coordination problems, players make use of a much wider range of properties of games than conventional theory treats as relevant. A simple example is a game described by Schelling in which each of two players, independently and without communication, has to write down either 'heads' or 'tails', each player's objective being to write down the same word as the other player does. In a conventional game-theoretic analysis, this is a completely symmetrical 2×2 game. But when the game is played in controlled experiments, substantial majorities of subjects choose 'heads'. As a result, those subjects (individually and collectively) achieve a degree of success that standard rational-choice theory cannot explain (Schelling, 1960, pp. 55–56). Schelling uses examples like this to argue that the conception of rationality used in that theory is inadequate:

> [T]o assert the influence of ... the symbolic and connotative details of the game ... does not involve the question of whether game theory is predictive or normative – concerned with generalizations about actual choice or the strategy of correct choice. The assertion here is *not* that people simply *are* affected by symbolic

> details but that they *should* be for the purposes of correct play. A normative theory must produce strategies that are at least as good as what people can do without them. More, it must not deny or expunge details that can demonstrably benefit two or more players and that the players, consequently, should not expunge or ignore in their mutual interest. (p. 98)

As this passage makes clear, Schelling sees his work as a contribution to normative game theory. But, for Schelling, normative game theory is ultimately a *practical* enquiry: the aim is to find principles that can assist real players to achieve their ends in real games with real co-players. It is a theory of correct play, but with 'correctness' defined in practical terms. I take it that Schelling is looking for principles of correct play which satisfy two criteria. Consider a real-world (two-player) game for which some putative principle of correct play makes recommendations to both players. The first criterion is that, for each player, it is in her interests to act on the recommendation, given what the other player can be expected to do *in fact*. The second criterion is that recommendations are not self-undermining. That is, it is in each player's interest to act on the recommendation if the other player acts on it too.[6] The first criterion highlights a significant aspect of Schelling's approach – that observed regularities in the behaviour of real game-players can be used as ingredients in a theory of rational play. Such a manoeuvre might be seen as illegitimate in an a priori theory of the interaction of ideally rational agents.[7]

I do not want to claim that Lewis endorses Schelling's pragmatic conception of rationality. (To the contrary, some of the distinctive features of Lewis's analysis of salience seem to reflect Lewis's desire to find rational-choice foundations for Schelling's theory.) Still, it is illuminating to ask what contribution Lewis's analysis makes to the enterprise that Schelling calls 'normative game theory'. To put the question more directly: What practical advice can Lewis give to agents who face coordination problems in real life? The answer, I submit, is that Lewis recommends to each agent the following presumptive principle of correct play: *Project regularities in population-wide behaviour, using the analogies that you perceive as most natural, then choose the action that can be expected to produce the best results for you, given that projection*. This principle is 'presumptive' in the sense that it asserts a presumption in favour of a certain conclusion, but can be overridden by other reasons if they exist and are sufficiently strong. Lewis's analysis shows why, in the absence of countervailing reasons, this principle is good advice to each individual and is not self-undermining.

It would be easy to supplement this advice by adding other presumptive principles. In particular, there are sometimes good reasons to override the presumption in favour of the most natural projection. Here is an example.

Suppose again that you and I are the Oberlin residents whose call (initiated by you) has just been cut off. Suppose that I am a new arrival in Oberlin. I have experienced only four previous instances of a cut-off phone call, each involving a different person. In all of them, I was the caller and I called back. In each case the connection was resumed immediately, suggesting that the person I called was waiting for me to call back. I perceive two ways of describing this regularity – 'In each case, I called back', and 'In each case, the caller called back'. If I project the first description, I will call you; if I project the second, I will wait for you to call me. The first description is more natural *for me* – it is more salient, it comes to my mind more immediately and more easily. But should I project it? If I ask myself which regularity is more likely to be embedded in an ongoing convention, there are reasons for favouring the second description.

One type of reason refers to the mechanisms by which conventions reproduce themselves, and so fits easily into Lewis's framework. For example, if there are many people in Oberlin, it is improbable that they are all using a rule which refers to me as a specific person. It is somewhat more likely that there is a rule which makes use of some potential asymmetry in personal characteristics, applicable to Oberlin people in general, and that there is a tendency for this rule to classify me as the person who should call back. (Perhaps the newer resident calls back.) But such a rule would not work very well unless the people of Oberlin were very well-informed about one another, while I already have some evidence that the actual rule works well (it has worked four times out of four). In contrast, the 'caller calls back' rule uses information that is readily available to, and psychologically salient to, all pairs of parties to phone calls.

Another type of reason refers to the mechanisms by which conventions emerge and decay, and by which the range of a convention expands or contracts. Lewis's analysis is not directly concerned with such mechanisms; but evolutionary game theory is, and it can provide some useful additional advice. One relevant finding of evolutionary game theory is that, other things being equal, less effective conventions are more vulnerable to destruction by the 'invasion' of rival rules. Thus, there is some presumption that behaviour in recurrent coordination problems will be governed by relatively effective conventions. That favours the 'caller calls back' rule. Another finding is that a regularity *R* is particularly well-placed

to become the actual convention if, in a situation in which no convention has yet emerged, there is some tendency for individuals to act *as if R* were the convention. For example, imagine a town in which, whenever a phone call is cut off, both parties immediately try to call back. The original caller is likely to have immediate access to the phone number of the person called, since he had to use it to make the call, while the person called will have more difficulty finding the caller's number. Thus, among those cases in which the connection is resumed smoothly, there are more instances of the caller calling back than of the called calling back. Before anyone becomes conscious of the 'caller calls back' rule, it is as it that rule were already being followed to some extent.

The upshot of all this is that Lewis's presumptive principle is not the last word in good advice to people who confront recurrent coordination problems. But it is presented only a presumptive principle, to be followed in the absence of stronger countervailing reasons. Understood in this way, it *is* good advice. In the example, if I (the new resident in Oberlin) recognise that my observations can be organised as instances of the 'caller calls back' regularity, it is probably sensible for me to project that description, even if 'I call back' is more natural to me. But that is because of the countervailing reasons I have explained. Suppose instead that the 'caller calls back' description does not occur to me, while the 'I call back' description does. Would I then be well advised to project the description that *does* occur to me? I maintain that the answer is 'Yes'. My four consecutive successes in coordination give me some reason to conjecture that, consciously or unconsciously, I have been following a rule that is well adapted to the behaviour of other Oberlin residents. If that is all I know, the best I can do is to try to continue using the same rule as before. If the only description I can find for what has happened is 'I call back', then that is the rule I should follow, until I make new observations, or become aware of new reasons. This remains true even if I think it unlikely (or even impossible) that the other Oberlin residents perceive the problem in terms of the description that I use. My four successes give me reason to conjecture that the rule they are using, whatever it may be, meshes with some regularity in my behaviour, and (by assumption) the only description I have for the latter is 'I call back'.

Of course, since these conjectures are themselves inductive inferences, the pragmatic arguments I have been rehearsing can make no impact on the *old* problem of induction. But, I maintain, they help us to see the credibility of Lewis's criterion of projectibility.

3. Breaking symmetries

Up to now, I have followed Lewis in focusing on how conventions reproduce themselves. I now consider the implications of his analysis of projectibility for the *origin* of conventions. More precisely, I will consider its implications for a particular theory of the origin of conventions, which uses the idea of *symmetry breaking*. That theory was first developed in theoretical biology, by John Maynard Smith and his collaborators (Maynard Smith and Price, 1973; Maynard Smith and Parker, 1976). I adapted the theory to apply to human conventions, replacing natural selection by experience-based learning (Sugden, 2004; first edition 1986). The same mechanism was independently rediscovered by Brian Skyrms (1996).[8]

The essential idea can be captured in the following simple model. Consider a single large population from which, recurrently, pairs of individuals are drawn at random to play a coordination game. In these games, players do not know (or cannot remember) one another's identities, with the implication that it is impossible for individual players to build reputations that can be carried over from one game to another. In each instance of the game, each of the two players chooses one of two actions, *call back* or *wait*. Payoffs are the same for all instances of the game and are completely symmetrical between players and between actions. If one player chooses *call back* and the other chooses *wait*, each gets one unit of payoff; otherwise, each gets zero.

In one version of the model, the players' positions are symmetrical in *all* respects, so that every game looks exactly like every other. Each player sees each game as an interaction between 'me' and 'another person'; no other differentiation between the two players is perceived. Thus, in any given game, all that one player knows about the other is that the latter has been drawn at random from the whole population. For any given time t, we can define the probability $p(t)$ with which an individual, randomly selected from the population, chooses *call back*. It is easy to see that if $p(t) < 0.5$, the expected payoff is higher for players who *call back* than for players who *wait*, while the converse if true if $p(t) > 0.5$. Thus, if individuals learn by experience to choose actions with higher expected payoffs, the system will gravitate towards a stable equilibrium at which *call back* is played with probability 0.5.

Now consider a more realistic version of the model. The actions and payoffs of the game are as before. In any given game, however, each player is aware of more than these formal properties. He also has a *view* of the game, which is to be interpreted as a description

of those contextual features of the game that the player recognises. Views may differ across players and across games. Thus, a player may use properties of views to organise her memories of past games, and her choice of action in a given game may be conditioned on her view.

I begin with a very simple example. Suppose that there are only two possible views, *caller* and *called*. In every game, one of the players, determined at random (independently for each game), views the game as *caller* while the other views it as *called*. For any time t, let $q(t)$ be the probability that a player, randomly selected from the population and given the view *caller*, chooses *call back*; and let $r(t)$ be the corresponding probability for a player whose view is *called*. If players ignore the information in their views, the analysis is just as before, and the system gravitates to a stable equilibrium at $q(t) = r(t) = 0.5$. But if players differentiate between games in which they are *caller* and game in which they are *called*, these two probabilities can evolve separately. For players who are *caller*, *call back* gives a higher expected payoff than *wait* if $r(t) < 0.5$, and the opposite if $r(t) > 0.5$. Symmetrically: for players who are *called*, *call back* gives a higher expected payoff than *wait* if $q(t) < 0.5$, and the opposite if $q(t) > 0.5$. If players learn by experience to choose actions with higher expected payoffs, and if they differentiate between their experiences as *caller* and as *called*, $q(t)$ and $r(t)$ will tend to evolve in opposite directions. (The greater the tendency for players with one view to choose one action – say, for *callers* to *call back* – the greater the incentive for players with the other view to choose the other action – that is, for the *called* to *wait*.) In consequence, the system has just two stable equilibria, one with $q(t) = 1$ and $r(t) = 0$ (*caller calls back, called waits*), the other with $q(t) = 0$ and $r(t) = 1$ (*caller waits*, *called calls back*). There is also an unstable ('knife-edge') equilibrium in which $q(t) = r(t) = 0.5$.

In the theoretical literature, this result is usually taken as showing that, in recurrent coordination problems in which coordination requires the two players to choose different actions, arbitrary asymmetries between players tend to precipitate the evolution of correspondingly arbitrary conventions. Even if the system begins at the knife-edge equilibrium, perhaps because at first no one thinks that apparently arbitrary differences between views could have any relevance to their decisions, random variation of some kind or another will sooner or later tip the system into the basin of attraction of one or other of the stable equilibria. Presenting this argument, Skyrms (1996: 73) says that, insofar as the aim is to show only that *some* convention will emerge, 'it doesn't really matter' what form the

random variation takes, but he suggests that there might be some slight variation in payoffs, or that players might have 'imperfect noisy memory'. My version of the argument was:

> Sooner or later, ... some slight asymmetry of behaviour will occur by chance; some players will think that something more than chance is involved, and expect the asymmetry to continue. Even though the expectation has no foundation, it is self-fulfilling. (Sugden, 2004: 45)

One problem with this argument is that 'sooner or later' may be a very long time. The process that leads to the emergence of a convention *R* begins when random variation produces a population-wide asymmetry in behaviour that mimics imperfect conformity to *R*. (For example: by chance, *callers* are more likely than *called* to *call back*.) Players observe this asymmetry and project it into the future. If, as in the model I have presented, there are only two putative conventions, players need relatively few observations of asymmetric behaviour in order to discover the form that that asymmetry takes. But if the information content of players' view is much richer than this – as it surely is in reality – there are many, many putative conventions. The more possible regularities that can be exhibited in the play of a given game (that is, a game with given specifications of actions and payoffs), the more observations are needed to identify any one regularity reliably (Sugden, 2004: 187–196). That is just another way of saying, with Lewis, that there are innumerable alternative analogies. Thus, an asymmetry in behaviour might have to continue for a long time before it was widely recognised. If the asymmetry was merely the product of transient random variation, it might disappear long before anyone recognised it.

The best answer to this objection, I think, is to appeal to salience. Among the innumerable set of alternative analogies, only a very small number are perceived as natural or salient by any given individual; and different individuals' perceptions of salience tend to show strong similarities. The pre-condition for the emergence of a convention is not the occurrence of just *any* population-wide asymmetry in behaviour: it is the occurrence of a population-wide asymmetry in behaviour that maps on to a *pre-existing* system of salient analogies.

For example, compare the following two alternative scenarios in which I am an Oberlin resident. In *Scenario 1*, I can remember the last ten instances of phone cut-offs involving me. They all occurred in the past month, and in all of them the other player's behaviour fitted the pattern 'caller calls back, called waits'. In fact (although I don't know this), 90 per cent of all cut-offs in Oberlin in the past month were resolved in this way. In

Scenario 2, I can remember the last ten instances, all in the past month, and in all of them the other player's behaviour fitted the pattern 'on Mondays, Thursdays and Fridays, caller calls back, called waits; on other days, caller waits, called calls back'. In fact (although I don't know this), 90 per cent of all cut-offs in Oberlin in the past month were resolved in this way. The difference between the two cases is that the regularity in Scenario 2, though no less real than that in Scenario 1, is perceived by me as less projectible. Thus, Scenario 1 is more likely than Scenario 2 to lead to the emergence of a convention. The implication is that conventions are precipitated by salient asymmetries. Or, to change the metaphor, the conventions that emerge carry the imprint of pre-existing conceptions of salience.[9]

This argument has an implication that may seem paradoxical: conventions evolve more easily if individuals are willing to project *accidental* regularities. Using the same simple model as before, suppose that each game is between a *caller* and a *called*, but that initially no one recognises this asymmetry. The system settles down in the symmetrical equilibrium in which $q(t) = r(t) = 0.5$. Then, players begin to become aware of the distinction between *caller* and *called*. Suppose that (as suggested by Skyrms) each player can remember only a small sample of her recent games. Then, as a result of sampling variation, a period t' may eventually occur in which most players' memories happen to contain a disproportionate number of games in which their opponents were *callers* who *called back*, or *called* who *waited*. If this accidental occurrence is to initiate the evolution of a convention, players must project the apparent regularity that is exhibited in the games they remember. But when this regularity is first projected, it is entirely accidental. For each individual, it is mere accident that her memory is skewed the way that it is. Further, it is mere accident that different players' memories are skewed the same way. It seems that if a player fully understood the sampling mechanism by which memories were formed, she would realise that skewness in her sample was evidence neither of any genuine pattern in other players' behaviour, nor of any systematic skewness in their samples. So does the symmetry-breaking argument depend on the assumption that players are acting like the casino gambler who bets on a particular roulette number because it won last time?

I shall try to cut through this paradox by arguing that Lewis's projectibilty criterion – the presumption that salient regularities are projectible – is pragmatically defensible, not just in the context of ongoing conventions, but as a general presumptive principle for learning about the world. In its general form, addressed to any agent, the principle is: *If you observe a regularity in the world, and if you perceive it as salient, treat it as projectible*. This

principle defines a demarcation line between valid and invalid projections: projections are valid for an agent to the extent that the projected regularities are salient for that agent. Thus, it is a potential solution to Goodman's Problem. At this point, my argument starts to have implications in the domain of philosophy of science.

4. Grounding inductive inferences on salience alone

As a starting point for explaining how the scope of Lewis's projectibility principle can be extended, I consider an example used by Goodman (1954). At one point, Goodman frames his Problem in terms of the confirmation of hypotheses. The idea here is that a hypothesis makes a general claim which applies to a class of 'instances'. If one of the implications of a hypothesis is found to be true, does that count as a confirmation of the hypothesis? To claim that it does is to make an inductive inference from an instance in which the hypothesis is known to be true to a class of unobserved instances. Goodman's answer to this question is that 'Confirmation of a hypothesis occurs only when an instance imparts to the hypothesis some credibility that is conveyed to other instances' (pp. 66–69). The crucial question to be resolved is then: 'What hypotheses are confirmed by their positive instances?' (p. 81). In trying to explicate this latter question, Goodman presents the following example:

> That a given piece of copper conducts electricity increases the credibility of statements asserting that other pieces of copper conduct electricity, and thus confirms the hypothesis that all copper conducts electricity. But the fact that a given man now in this room is a third son does not increase the credibility of statements asserting that other men now in this room are third sons, and so does not confirm the hypothesis that all men now in this room are third sons. ... The difference is that in the former case the hypothesis is a *lawlike* statement; while in the latter case, the hypothesis is a merely contingent or accidental generality. Only a statement that is *lawlike* – regardless of its truth or falsity or its scientific importance – is capable of receiving confirmation from an instance of it; accidental statements are not. Plainly, then, we must look for a way of distinguishing lawlike from accidental statements. (p. 73)

The suggestion here seems to be that the proposition 'All men now in this room are third sons' (applied to the room in the University of London in which Goodman is delivering his second Special Lecture in Philosophy) can be ruled out as a valid projection, *prior to any observations*. However many third sons we may happen to find in the room, that will be 'merely contingent or accidental', and therefore uninformative about other men in the room. In more abstract terms: the problem of deciding what regularities are potentially projectible (that is, what regularities should be projected, were they to occur) can be re-cast as a

problem of discriminating between 'law-like' and 'accidental' *statements*; this discrimination should be made by reference to formal or linguistic properties of the statements themselves, independently of their truth values. I disagree.

Suppose I am a woman in the lecture room, just before Goodman starts to speak (and so before he has introduced the idea of third sons). I count the number of men in the room; there are 50. I pick one man and ask him his position in the birth order of his mother's sons. He says he is a third son. This does not strike me as anything out of the ordinary, even though third sons are relatively rare (my rough guess is that about one man in 14 is a third son).[10] I treat it merely as a fact about that particular man, telling me nothing about other men in the room. I pick a second man and ask him the same question. He is a third son too. That strikes me merely as a coincidence. But what if I have asked five men, and they are all third sons? Now, there is no question that the data generated by my enquiries has a salient pattern, namely that all the men I have questioned so far are third sons. Perhaps I still interpret this pattern as a coincidence, but the coincidence is so striking that it forces itself on my attention. What if the next five men are third sons too? The probability of observing ten third sons in a row is of the order of 1 in 250 billion. If I exclude the first five cases on the grounds that it was only after observing those cases that third sons became uniquely salient to me, the probability of observing even five third sons in a row is about 1 in 500,000. That is not the kind of coincidence in which I am prepared to believe. I conclude: *the pattern is not accidental*. Or: *something is going on here*.

What these expressions imply is the conviction that *some* mechanism is at work, systematically creating the pattern that has been observed. And that in turn implies the presumption that that pattern is projectible within *some* domain, namely the domain in which the mechanism operates. I may have very little idea about what the mechanism is, or about the domain in which it operates, in which case I will be unsure about how far the pattern can be projected, or in which directions. But even so, I can feel reasonably confident in projecting it to cases that I perceive as sufficiently similar to those I have already observed. In the case of the lecture, if I think that the procedure by which I picked out the first ten men was not too far from random sampling from among the 50 men in the room, I will infer that if I use the same procedure to pick more of those men, each of the men I pick will very probably be a third son. Such inferences, I submit, are pragmatically justified – that is, our experience tells us that they tend to be confirmed.

The mode of reasoning I have been describing can be represented by the following *Salience Schema*:

I have observed regularity *R* in domain *D*.

I perceive *R* as salient.

If events in *D* were governed by chance, salient regularities as extreme as *R* would occur with very low frequency.

Very probably, *R* is not an accident.

The 'very low frequency' in the third premise is to be understood as an objective statement about the properties of some random mechanism. In contrast, the 'very probably' in the conclusion is to be understood subjectively, as expressing a degree of belief on the part of the agent doing the reasoning. Take the case of the lecture. Here *D* is the set of ten men I have sampled and *R* is their all being third sons. That 'events in *D* are governed by chance' is a null hypothesis defined in terms of a specific random mechanism, say that each man is drawn at random from a population in which the distribution of birth orders is the same as for British men as a whole. The relative frequency with which ten third sons will be found in a sample of ten drawn from that population is a statistical fact. But the fact that that relative frequency is low does not entail, as a matter of logical necessity, that the agent has a high degree of belief that the observation of the ten third sons is non-accidental. The Salience Schema is a schema of inductive inference, not of deductive logic.

Notice that for 'Very probably, *R* is not an accident' to be derived as a conclusion of the Salience Schema, it is not enough that *R* is very unlikely to occur by chance. *R* must be salient, and it must be very unlikely that *any* salient regularity would occur by chance. For example, consider the following regularity *R*′: 'The first ten men sampled, in order of sampling, are first, third, first, first, second, second, first, fifth, first and first sons'. Given the same null hypothesis as before, the objective probability of finding exactly this pattern is of the order of 1 in 200,000. But *R*′ is not a salient pattern. Every possible sequence of ten birth orders has *some* pattern, and every such sequence has a very low objective probability – simply because there are so many possible sequences. The special feature of *R* is that it happens to coincide with one of the very few patterns that I perceive as salient.

Perceiving a pattern as salient is not the same thing as having a prior expectation of finding it in the world.[11] Like Goodman, I have no prior expectation of finding third sons grossly over-represented in philosophy lectures. The crucial mechanism is one of *surprise*. It is because finding salient birth-order patterns in lecture audiences is so extremely unlikely, given my background expectations, that I feel compelled to question those expectations when such a pattern appears. Thus, we should not require that a criterion of projectibility is grounded in beliefs about what kinds of patterns are likely to be found in the world.

This is important for understanding what might seem a strange property of the Salience Schema – that an agent's personal perceptions of salience are relevant in determining whether, for her, a particular regularity is projectible. If the criterion of projectibility was interpreted as expressing prior beliefs about the world, it would be natural to ask why the agent feels entitled to assume that her personal perceptions of salience map onto so-far unobserved regularities in the world. But the Salience Schema rests on no such assumption. As far as the logic of inference is concerned, the special significance of salient patterns is that merely that there are very few of them. Each person is qualified to check whether this is true for her.

This might seem to open the way for a different objection to the Salience Schema. Reasoning according to that schema cannot identify a regularity as non-accidental unless the agent's perceptions of salience happen to map onto it. If those perceptions have no privileged correspondence with the as-yet unknown truths of the world, isn't such reasoning likely to be ineffective in discovering those truths? My answer is that we often discover scientific truth indirectly, by first noticing regularities that happen to be salient *to us*, and only later finding that these are the products of general mechanisms, most of whose effects we had been unaware of. For the first step to be possible, all that is necessary is that *some* effect of the relevant mechanism is sufficiently salient. For example, a crucial step in the development of Newtonian physics was Kepler's discovery that planetary orbits are elliptical. What, one might ask, is so special about an ellipse? Perhaps the answer is not that the ellipse is part of the deep causal structure of the universe, but merely that it happened to be a salient shape for seventeenth-century scientists. But Kepler's discoveries concern only a tiny part of the body of regularities that Newtonian physics explains; scientists with different perceptions of salience might have found the same physics by a different route.

Goodman's analysis seems to me to concede too much to the thought that projectibility is about prior expectations. His solution to his Problem discriminates between

predicates according to the degree to which they are *entrenched* in the linguistic practices which accompany (or, he might prefer to say, constitute) inductive inference. The more entrenched a predicate, the more projectible are statements which refer to it. Or, in an alternative wording, the judgement that a predicate is projectible derives from its 'habitual projection' (p. 98). Thus, in terms of Goodman's central example, '"green", as a veteran of earlier and many more projections than "grue", has the more impressive biography', and so is defined to be more entrenched (p. 94). He sums up his proposal by declaring that 'the line between valid and invalid predictions (or inductions or projections) is drawn upon the basis of how the world is and has been described and anticipated in words' (p. 121). Notice the hint that the language we habitually use in inductive inference *anticipates* how the world really is. And the idea that projectibility depends on linguistically encoded habits *of projection*, rather than on linguistic practices in general, surely draws some of its intuitive appeal from the thought (itself an inductive projection) that, other things being equal, predicates that have been used in more previous projections are more likely to prove reliable in future projections. The Salience Schema appeals to the very different intuition that among the set of putative regularities, so few of them are salient that salient regularities are very unlikely to appear by accident.

I submit that the Salience Schema represents a mode of reasoning that is widely used in the natural and social sciences. It corresponds with the methods of classical (as contrasted with Bayesian) statistics, in which knowledge is accumulated by the rejection of null hypotheses. A substantive 'alternative' hypothesis about the properties of an unobserved population is accepted if observations of a sample show a pattern that is consistent with that hypothesis and if that pattern would be very unlikely to occur, were the null hypothesis true.[12] In order for such a test of statistical significance to be informative, the alternative hypothesis must have some prior salience, rather than being constructed ex post as a description of the evidence.

This is not to say that the alternative hypothesis must have been *predicted* in advance. Many important scientific developments have begun with the discovery of surprising and initially unexplained regularities in the world. Scientists have judged these regularities to be too surprising to be accidental, and have responded by looking for the missing explanations. Surprise requires a prior conception of salience, but that need not involve any prediction.

For example, an extremely productive programme of empirical and theoretical research in epidemiology was sparked off when, from the late 1970s, a large-scale longitudinal study of the health of British civil servants began to reveal that mortality rates have a steep 'social gradient', declining progressively as one moves up the civil service hierarchy.[13] The British civil service has a strongly hierarchical structure, with four clearly defined grades with a rank order from 'office support' at the bottom to 'administrative' at the top. Thus, cross-grade differences in mortality, once observed, are highly salient for anyone who works with civil service data. However, since there were no obvious health-related differences in working conditions between the grades, the researchers were at first very surprised at their findings. Subsequent research has shown that, in many human (and other primate) societies, there is a systematic negative association between social status and mortality, and evidence is beginning to accumulate which suggests a causal mechanism based on stress.

Another example is discussed by Ronald Giere (1988, pp. 227–277). From the 1920s to the early 1960s, the hypothesis of continental drift, first proposed by Alfred Wegener, was controversial. For its proponents, this hypothesis was the only credible way of accounting for an astonishing body of evidence of geological, paleontological and biological correspondences between widely-separated continents. However, as its opponents pointed out, there seemed to be no physical mechanism which could move continental masses such enormous distances. The geologist Drummond Matthews later recounted how he had been convinced of the truth of the hypothesis while on a mapping expedition to the Antarctic. He had expressed scepticism to a fellow geologist, Raymond Adie:

> [Adie] said, 'Oh well, ... if you don't believe in continental drift just take a tape measure and measure the Devonian sections in the Falkland Islands.' I did, and they were very much impressively the same as the description [of geological sections in South Africa].... Inch by inch they measured up. So, I came back quite enthusiastic. (quoted by Giere, p. 253)

Adie's argument has the structure of the Salience Schema. A bed of sedimentary rock can be described by the order in which different materials have been deposited and by the depth of each layer of deposits. Two beds of rock, one in the Falkland Islands and one in South Africa, are found to have almost exactly the same characteristics. Although there is an infinity of possible patterns of correspondence between two beds of rock, the pattern that geologists would count as 'being the same' is highly salient. The objective probability that two beds of rock would correspond in this particular way, had they been laid down in widely

separated continents, is infinitesimally small. Even if we have no idea about how this regularity could possibly have been caused, we can be as sure as is humanly possible that it is not an accident.

5. Conclusion

I conclude that inductive inferences can be grounded on salience alone. More precisely: an agent's own perception of salience provides her with a criterion of projectibility with which she can overcome Goodman's Problem. If that is right, an apparent paradox in the theory of conventions is dissolved. The paradox is that, in explaining the emergence of conventions, the theory assumes that individuals project salient regularities in other people's behaviour even though those regularities are entirely accidental. Provided that conceptions of salience are shared within the relevant population, each person's projection leads him to behave in a way that confirms other people's projections; but (it might be said) we lack an explanation of why any of these projections are made in the first place. The answer I have given in this paper is that, in order for a regularity to be projectible, it is sufficient that it is salient and that the probability of the occurrence of a salient regularity is sufficiently low. But even highly improbable events sometimes occur. On those rare occasions, human agents project accidental regularities – and are pragmatically justified in doing so.

References

Cubitt, Robin and Robert Sugden (2003). Common knowledge, salience and convention: a reconstruction of David Lewis's game theory. *Economics and Philosophy* 19: 175–210.

Giere, Ronald (1988). *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

Gilbert, Margaret (1989). Rationality and salience. *Philosophical Studies* 57: 61–77.

Goodman, Nelson (1954). *Fact, Fiction and Forecast*. Cambridge, Mass.: Harvard University Press. Page references to fourth edition, 1983.

Hume, David (1739–40 / 1978). *A Treatise of Human Nature*. Oxford: Oxford University Press.

Lewis, David (1969). *Convention: A Philosophical Study*. Cambridge, Mass.: Harvard University Press.

Marmot, Michael (2004) *Status Syndrome: How Your Social Standing Directly Affects Your Health and Life Expectancy*. London: Bloomsbury.

Maynard Smith, John and Geoffrey Parker (1976). The logic of asymmetric contests. *Animal Behaviour* 24: 159-175.

Maynard Smith, John and George Price (1973). The logic of animal conflicts. *Nature, London* 246: 15-18.

Mayo, Deborah (1996). *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Schelling, Thomas (1960). *The Strategy of Conflict*. Cambridge, Mass.: Harvard University Press.

Skyrms, Brian (1996). *Evolution of the Social Contract*. Cambridge: Cambridge University Press.

Sugden, Robert (1998). The role of inductive reasoning in the evolution of conventions. *Law and Philosophy* 17: 377–410.

Sugden, Robert (2004). *The Economics of Rights, Cooperation and Welfare*, second edition. Basingstoke: Palgrave Macmillan. First edition 1986.

Sugden, Robert (2006). Hume's non-instrumental and non-propositional decision theory. *Economics and Philosophy* 22: 365–391.

Sugden, Robert and Zamarrón, Ignacio (2006). Finding the key: the riddle of focal points. *Journal of Economic Psychology* 27: 609-621.

Notes

[1] I defend these readings of the *Treatise* in Sugden (2004, 2006).

[2] For more on Lewis's contributions to game theory, see Cubitt and Sugden (2003). My own work on conventions, which was one of the earliest applications of evolutionary game theory in economics, was influenced by all three of Hume, Schelling and Lewis (Sugden, 2004).

[3] Lewis's definition of 'coordination equilibrium' requires in addition that, for each agent, it he performs his component of the pair, he prefers that the other performs hers. This condition is significant for Lewis's analysis of norms, but is not necessary for my purposes.

[4] Lewis's analysis seems to have been the first attempt to explain focal points (i.e. the phenomenon discovered by Schelling) using the formal tools of conventional game theory. This approach has subsequently been followed by a number of game theorists. Schelling's own explanation of focal points may be different. On this, see Sugden and Zamarrón, 2006.

[5] Recall that in his treatment of 'coordination by salience', Lewis uses a slightly different formula for what presumably is intended as the same idea: 'when [we] have no stronger ground for choice'. As Gilbert (1989) points out, the latter formula can be read in a way which makes Lewis's argument self-refuting. If the argument claims to prove that it is rational to choose the salient, it can succeed only by contradicting the supposition that the salient is picked *in the absence of stronger reasons* than salience itself. However, the supposition that *there are no reasons to do otherwise* is not undermined.

[6] In stating these conditions, I presuppose that players have 'interests' that are adequately represented as payoffs of the game.

[7] For more on this pragmatic reading of Schelling, see Sugden and Zamarrón, 2006.

[8] In my account, this mechanism works in conjunction with salience in the Schelling–Lewis sense. Skyrms argues that, in an evolutionary theory, salience 'is no longer required' (p. 102). I disagree, for reasons explained later in this paper.

[9] This mechanism is discussed in more detail by Cubitt and Sugden (2003).

[10] My back-of-the-envelope calculations are based on the assumption that any mother's children are equally likely to be boys or girls. Suppose that 20 per cent of mothers have exactly one child, 40 per cent have two, 25 per cent have three, 15 per cent have four, and 5 per cent have five (giving an average of 2.4 children per mother). Then the proportions of

men who are first, second, third, fourth and fifth sons will be 0.634, 0.279, 0.073, 0.013 and 0.001 respectively.

[11] A reader who is committed to Bayesian rationality may object that an agent who reasons according to the Salience Schema must have started out with a prior subjective probability for *R* that is greater than its objective probability according to the null hypothesis. That is true *for an agent whose beliefs conform to the axioms of Bayesian rationality*. But, I suggest, when we think about the role of salience in inductive reasoning, the most natural conclusion is that normal human reasoning is not entirely Bayesian. In particular, surprising evidence can create completely new beliefs, rather than merely inducing an agent to update pre-existing priors.

[12] For an account of scientific knowledge in which thus kind of reasoning is central, see Mayo (1996).

[13] This research programme was led by Michael Marmot. For a popular account of this work, see Marmot (2004).