Salience as an emergent property

In the classical theory of noncooperative games, the formal representation of a game in normal or strategic form takes no account of how players and strategies are labelled. Arbitrary player labels (such as 'row player' and 'column player') and strategy labels (such as 'up' and 'down') may be used by the theorist as an aid to analysis, but these are not intended to represent how the players describe the game to themselves, and are not treated as part of the specification of the game; solution concepts are defined so that they are independent of such labels. However, it is widely recognised that in real-world games, players often *do* take account of the labels that feature in their own descriptions of those games, and that these labels can play an important role in equilibrium selection. As first hypothesised by Thomas Schelling (1960), players recognise (and expect their co-players to recognise) that, by virtue of differences in labelling, some equilibria are more 'prominent' or 'salient' than others; there is a systematic tendency for the most salient equilibrium (the 'focal point') to be selected.[1] There is now a large body of experimental evidence which confirms this hypothesis in relation to pure coordination games (e.g. Mehta et al., 1994a, 1994b; Bacharach and Bernasconi, 1997; Crawford et al., 2008; Bardsley et al., 2010), and some evidence that this result extends to tacit bargaining games that are framed in ways suggestive of real-world bargaining (Isoni et al., 2011). But although there have been a number of attempts to incorporate labels into formal game theory (e.g. Lewis, 1969; Gauthier, 1975; Sugden, 1995; Bacharach and Stahl, 2000; Casajus, 2001; Janssen, 2001; Bacharach, 2006), theorists have had little success in finding general characterisations of the features that make some labels more salient than others. For this reason, the concept of salience has never been truly assimilated to game theory, even though most game theorists acknowledge the truth of Schelling's insights.

In explaining the concept of salience, Schelling (1960) relies heavily on metaphors. He describes players as searching for a 'clue', 'key', 'hint', 'message', 'signal' or 'suggestion' that is hidden in their decision problem. He says that a focal point is 'prominent', 'conspicuous' or 'obvious'; it has a 'claim to attention', a 'power of suggestion'

[1] In this paper, we will use the term 'salient' to refer only to properties of labels and to similarity relationships between labels.

or an 'intrinsic magnetism'; it 'commands recognition', 'suggests itself', exerts 'discipline', 'communicates its own inevitability', or 'dictates' that it is to be chosen.[2] He emphasises that focal points can be found by many different methods, including the use of 'analogy', 'precedent', 'aesthetic or geometric configuration', 'casuistic reasoning', and 'whimsy' (p. 57). He makes much use of illustrative examples, which are intended to help the reader to understand the common features of focal points. For example, in a coordination problem in which the players are trying to give the same answer to 'Name any positive number', the focal point is '1'; to 'Name a place in New York City to meet the other person', the focal point (for people in New Haven, Connecticut in the 1950s) is 'Grand Central Station'; to 'Name a time of day to meet the other person', the focal point is '12 noon'. What is salient for any specific pair of players may depend on their common experiences and cultural repertoires, but the *concept* of salience is to be understood as a generic feature of games.

For Schelling, it seems, finding focal points is more of an art than a science. The skills that are required can be taught and learned, but seem incapable of being reduced to mechanistic rules. Thus, game theorists who have read *Strategy of Conflict* can recognise focal points when they see them, but cannot represent Schelling's ideas in the formal language of their theory.

This is not to deny that specific aspects of salience can sometimes be modelled by means of relatively minor additions to classical game theory. For example, the intuitive idea that uniqueness confers salience can be represented in theoretical models in which the set of possible labels is highly restricted (Casajus, 2001; Janssen, 2001; Bacharach, 2006). Similarly, the intuitive idea that frequent public references to an object confer salience can be represented in a model in which players observe independent samples of a common pool of references (Sugden, 1995). But if one tries to represent the feature that is common to the salience of '1' among the set of positive numbers *and* to the salience of 'Grand Central Station' among the set of potential meeting places in New York, game-theoretic modelling seems to be of little use. Nevertheless, there clearly *is* a common feature, which Schelling is able to describe by metaphor and example. As we have pointed out, the experimental evidence shows that individuals *do* tend to choose salient strategies in coordination games. So salience is not just a concept that ordinary people are capable of learning: it is a concept

[2] For page references to Schelling's many uses of these metaphors, see Sugden and Zamarrón (2006).

that they already perceive and use as a means of solving coordination problems. It is therefore natural to ask how common perceptions of salience originate.

In this paper, we sketch a possible answer to this question. We consider the mechanisms by which, when the members of a population face recurrent coordination problems, conventions of behaviour emerge. We argue that labelling plays an important role in these mechanisms. In consequence, the conventions that emerge in different areas of social life tend to have certain common labelling characteristics. By becoming aware of these characteristics, whether consciously or unconsciously, individuals become more proficient in recognising the conventions that other people are following and in anticipating the conventions they will find in unfamiliar settings. We suggest that these common characteristics are significant components of the concept that game theorists call 'salience'. On this account, salience is an emergent property of human interaction.

In Section 1, we explain the sense in which we are using the concept of 'emergence'. This understanding of emergence is closely related to the idea of spontaneous order, as analysed by Friedrich Hayek. In some significant respects, the same understanding is implicit in modern evolutionary game-theoretic analyses of the emergence of conventions in recurrent play of coordination games. However, most evolutionary game theory has taken no account of how the emergence of conventions might be influenced by properties of labelling. We argue that an adequate theory of experiential learning needs to take account of labelling.

In the remainder of the paper, we present a simple model of salience as an emergent property of recurrent interaction, and discuss some relevant experimental evidence. Our model is based on two ideas. The first is that one of the mechanisms by which experiential learning works is a tendency for an individual, when facing a new decision problem, to recall previous problems that are perceived as similar to it and to choose an action that is perceived as similar to actions that, when chosen in those problems, were followed by favourable outcomes. This mechanism, applied in the context of games against nature, has been modelled by Itzhak Gilboa and David Schmeidler (1995) as 'case-based decision theory'. The underlying idea is common to many theories of inductive learning; indeed, Gilboa and Schmeidler cite David Hume's (1739–40/ 1987) theory of induction as an inspiration. The second idea is that, in the context of recurrent games, perceptions of similarity can be based on labelling, and that some kinds of labelling similarity are more reliably perceived than others. Thus, a putative convention is more likely to emerge and reproduce itself, the more

capable it is of being described in terms of reliably-perceived similarities. The intuition that some features of the labelling of strategies are more readily perceived than others, and so are more likely to seed conventions, has been expressed before (e.g. Schlicht, 1988; Sugden, 2004), but we believe that our analysis in terms of similarity relations is new.[3]

In Section 2, we offer an intuitive account of how the emergence of conventions may be the product of similarity-based learning. In Section 3, we present a formal model of this learning mechanism in a very simple environment, in which pure coordination games of the kind studied by Schelling are played recurrently. Deliberately, we model a version of the mechanism that makes minimal demands on individuals' cognition and memory. In acting in accordance with this mechanism, individuals do not engage in any kind of strategic reasoning; they simply attempt to repeat actions that have been successful in the immediate past. We show that, even in a population of such low-rationality agents, there is a tendency for salience-based conventions to emerge, 'salience' being defined in terms of reliable perceptions of similarity. In Section 4, we discuss evidence from an experimental investigation of recurrent play of coordination games. This evidence suggests that pairs of co-players learn to use similarity-based rules which increase the success with which they coordinate, and that the process of learning is based on the replication of co-players' previous choices.

1. Emergent properties and spontaneous order

The concept of 'emergence' can be used in many different ways. In this paper, we are concerned with emergent properties that are characteristic of *spontaneous order*, broadly in the sense in which Hayek (1973: 35–54) uses that expression. For our purposes, it is most useful to think of spontaneous order not as a property of reality, but as a property of theories or models that purport to explain or represent reality.

A spontaneous order is a pattern that is created when a population of entities interact together. Each member of the population acts on its own 'laws of motion', but their

[3] The general hypothesis of reinforcement learning – that is, the tendency to repeat rules or actions that have been successful in the past – is frequently used to explain learning in repeated games (e.g. Roth and Erev, 1995). There has been some work on how such learning can be transferred between games that, *strategically*, are similar but not identical; this can lead to the emergence of conventions that specify particular solution concepts such as payoff dominance (e.g. Rankin et al., 2000). In contrast, our paper is concerned with similarity relations between *labels*.

interactions create coherent patterns at the population level that are not simple aggregates of properties that the individual entities possess in isolation. Patterns that are created by such self-organising processes are *emergent properties* of interaction.

Consider an example from natural science – the properties of scree slopes. Scree slopes are made up of countless numbers of loose rocks, lying against the sides of mountains. These slopes are strikingly regular, with an even gradient from top to bottom. This regularity can be explained as a result of a simple property of loose rocks. For a given type of rock, there is a critical angle of slope; if an individual rock is placed on a slope steeper than this, it will tend to roll down, while if it is placed on a slope less steep than this, it will not move unless it is hit by some other body. Because of erosion, pieces of rock are constantly breaking loose from mountain sides and rolling downwards. If any area of a scree slope is temporarily steeper than the critical angle, it is unstable, and rocks tend to roll away from it; if any area is less steep than the critical angle, it tends to accumulate rocks which roll down from above and then stop. There is thus a constant tendency for the gradient of the slope to adjust itself towards the critical angle. According to this explanation, the regular gradient of scree slopes is a spontaneous order.

Notice that a general spatial property of large masses of rocks is being explained without any detailed analysis of the motions of individual rocks as they fall from above, move downhill and collide with other rocks. However, there is a presumption that those motions are fully explained by the principles of physics; and no additional causal factor is invoked in explaining the gradient of the scree slope. In this sense, the properties of the slope *supervene* on the properties of the rocks. Crucially, the pattern at the population level is created through the *interaction* of the individual components in a process with feedback loops. In contrast, the mere accumulation of masses of fallen rocks below mountains is not a spontaneous order, even though it is a predictable population-level consequence of the laws of motion of the individual rocks. That consequence is a simple aggregation of the tendency of individual rocks to fall, while the regular gradient of a scree slope is an emergent property of interaction among rocks.

As an example of spontaneous order in human interaction, consider Schelling's (1978: 137–166) 'checkerboard' model of racial segregation. (Recall that we are treating spontaneous order as a property of models, not of reality itself. Whether or not Schelling's model provides a good explanation of segregation in twentieth-century American cities, there

is a clear sense in which it describes a spontaneous order.) In this model, individuals are divided into two racial groups; each individual prefers not to live in neighbourhoods where there is a large majority of the other group. Individual choices based on these preferences interact to induce extreme segregation, even though no one wants this: segregation is an emergent property of interaction.

Until relatively recently, most game theory did not use spontaneous-order modes of explanation. The Nash equilibrium 'solutions' that game theory investigated were assumed to be common knowledge among perfectly rational players, either because those solutions were supposed to be accessible to rational agents by some unmodelled process of reasoning, or because they were 'suggested' to players by some unmodelled external authority. Game theorists hoped to narrow down the set of 'reasonable' equilibria by abstract analysis (e.g. Harsanyi and Selten, 1988). Considered in this conceptual framework, focal points tended to be seen as anomalous. Perfectly rational players, it was thought, would not need to base their decisions on salience – or would not be able to justify doing so.

There was a fundamental shift in the methodology of game theory in the late 1980s and early 1990s, when evolutionary analysis began to be widely used. Evolutionary game theory studies games that are played recurrently by individuals drawn from large populations. Individuals are not assumed to be perfectly rational, but merely to gravitate towards utility-maximising choices in dynamic processes of experiential learning. Equilibrium concepts are useful to the extent that they identify stable stationary points in such processes. Notice that stable equilibrium is a population-level property that is generated through a process of interaction between individuals, and that (because what is utility-maximising for one individual depends on how other individuals behave) feedback plays an essential role in this process. In other words, stable equilibrium in an evolutionary model is a case of spontaneous order. Expressing the same idea the other way round, Schelling's checkerboard model was evolutionary game theory *avant la lettre*.

Evolutionary game theory offers an explanation of social conventions as emergent properties of recurrent interactions between individuals, each of whom is pursuing his own interests given his expectations about other individuals' behaviour, where those expectations are grounded in his previous experience (e.g. Skyrms, 1996; Young, 1998; Sugden, 2004). Familiar examples of conventions that can be explained in this way include traffic rules, rules

for resolving conflicts over valuable resources, and common expectations among traders about the medium of exchange.

But despite this methodological shift, salience still receives little attention from game theorists.[4] In standard evolutionary game-theoretic models, learning processes are represented in models (for example, replicator dynamics or fictitious play) which make no reference to how strategies are labelled. If players' strategies and payoffs are completely symmetrical with one another, conventions are modelled as originating in symmetry-breaking perturbations. The thought seems to be that when games are played recurrently, salience is a redundant explanatory concept.

But as Robin Cubitt and Sugden have argued, that thought is mistaken (Cubitt and Sugden, 2003; Sugden, 2004, 2011). Standard evolutionary accounts of learning fail to deal with the following problem.[5] Experiential learning requires inductive inferences which project perceived regularities in a person's experience of previous games. Since no two games are exactly alike, those projections must rely on perceptions of similarity between non-identical games. Such perceptions are subjective, and so are likely to be sensitive to labelling. In the recurrent real-world interactions that evolutionary game theory is attempting to represent, a person's past observations can typically be fitted to a vast number of different logically possible patterns of similarity, with respectively different projections onto new games. Inductive inference works because only a small number of these patterns are recognised and perceived as 'natural' or 'obvious' . The only regularities that have the potential to reproduce themselves as conventions are those that fit pre-existing perceptions of obviousness that are shared by members of the relevant population. If this conclusion is correct, shared perceptions of similarity play an essential role in experiential learning. In consequence, one of the patterns to be found in the spontaneous order of conventions may be a tendency for conventions to be aligned with shared perceptions of similarity.

So far, our discussion has been very abstract. To help explain our ideas, we now consider a concrete example of how the emergence of a convention might be influenced by perceptions of similarity.

---

[4] One exception is Binmore and Samuelson's (2006) evolutionary analysis of games in which it is costly to 'monitor' labels. Since this model of the evolution of salience is very different from ours, we do not discuss it further.

[5] This problem seems to have been first noticed by Lewis (1969). See also Goyal and Janssen (1996).

2. The Right Turn Problem

At most British road junctions, driving behaviour is governed by priority rules that are designated by the highway authorities. Normally, one road through a junction is assigned priority; 'Give way' signs and road markings on the other routes signify that vehicles on those routes must yield priority to the traffic on the major road. This system works well at T-junctions, but has proved to be dangerous at crossroads – so much so that, outside towns, crossroads have progressively been replaced by pairs of slightly offset T-junctions. Where crossroads survive, drivers continually confront what we will call the Right Turn Problem. Consider a crossroads at which the east-west road has priority. Suppose this road is clear. One vehicle is approaching the junction on the minor road from the north, indicating the intention to go straight ahead; another is approaching on the minor road from the south, signalling to turn right. Since Britain drives on the left, their paths will cross. Which vehicle should give way to which? Failure to resolve this coordination problem can result in two stationary vehicles in the middle of the junction – a very dangerous outcome, since major-road drivers, who do not expect to have to give way to anyone, may be approaching the junction at speed.

Surprisingly, the Highway Code (the official codification of the rules of the road in Britain) provides no guidance on this question. If one thinks of the minor road as a continuous route across the junction, it may seem natural to give priority to the vehicle going straight ahead; but that perception is weakened by the significance that has been assigned to the major road. An opposing thought, encouraged by experience of offset T-junctions, is that the right-turning vehicle is joining the major road, and so inherits the priority of major-road vehicles after it has turned. But these opposing priority rules are not the only ways of resolving the Right Turn Problem. Another possible rule (and the one that seems to be most commonly used in practice) is to give priority to whichever of the minor-road vehicles reached the junction first; if there are queues, a vehicle is deemed to have reached the junction only when it gets to the front of its queue. Or priority might be given to the larger vehicle – or to the smaller. And so on.

If every instance of the Right Turn Problem were *exactly* the same as every other, this multiplicity of possible rules would not be an obstacle to coordination. For example, suppose that in every case the first driver to reach the junction were also the one going straight ahead.

Then ‘Straight ahead has priority’ and ‘First arrival has priority’ would merely be different ways of describing the same convention. For the purposes of an evolutionary analysis of the emergence of conventions, it would not matter which of these descriptions drivers used when thinking about the game, or whether some used one description and some the other. In reality, however, different instances of the Right Turn Problem are *not* identical. In every case there is a straight-ahead driver and a turning driver, and in every case there is a first arrival and a second arrival. But in some cases the straight-ahead driver is also the first arrival, and in other cases she is not. Because of this lack of alignment of the two ways of describing the asymmetry in the game, the descriptions that players use *do* matter. If the members of a population of drivers are to learn to coordinate their behaviour, they need to learn to use a common description of the game.

In this learning process, subjective perceptions of similarity may play a crucial role. Imagine you are a driver facing a new instance of the Right Turn Problem. You were the second driver to arrive at the junction and you are turning right. You recall just one previous instance of the problem. In that case, you were the first arrival and you were turning right, and the other driver clearly gave way to you. Does that recollection induce some expectation that the new driver will give you priority (since you are turning right) or that she will assume priority herself (since she was the first arrival)? The answer seems to depend on which similarity relationships are more salient for you, in the sense of coming more immediately to mind when you compare one case with another. If most people perceive ‘first arrival’ to be a more salient dimension of similarity than ‘turning right’, then one might expect a ‘first arrival’ convention to be more likely to emerge.

A related issue is the interpersonal reliability of judgements with respect to given dimensions of similarity. Suppose, for example, that when two vehicles arrive at a junction at approximately the same time, judgements about which was the first arrival are subject to noise. Then, even if both drivers follow the rule of giving priority to the perceived first arrival, they will not necessarily coordinate. One might expect that effect to tend to work against the emergence of the ‘first arrival’ convention. Taking another example, consider the rule of giving priority to the larger vehicle. If perceptions of the largeness of vehicles differ between individuals, two drivers who follow this rule may fail to coordinate. Notice that what ultimately matters is not whether individuals’ judgements are *correct* in relation to some objective criterion (such as the time at which a vehicle reaches a junction, or the volume or

mass of a vehicle), but whether they *coincide*. The interpersonal reliability of similarity judgements depends on shared subjective perceptions.

We conclude that an adequate analysis of the emergence of conventions needs to take account of the relative salience and reliability of different conceptions of similarity that can be applied to the labelling of games. This kind of analysis may seem to involve a major departure from mainstream game theory, but we will try to persuade the reader that the problems that have to be solved are theoretically tractable and amenable to experimental investigation.

3. A model of learning in recurrently similar games

In this section, we develop a model of learning behaviour in a family of *recurrently similar* games – that is, games that are similar but not identical to one another and that are played recurrently in a population. We assume a population that is sufficiently large to legitimate the use of the law of large numbers. In each *period* $t = 1, 2, ...,$ pairs of individuals are drawn at random from this population to interact as co-players. Each individual participates in one such interaction in each period.

Each interaction is a pure coordination game, defined by a set of n *labels*, where $n \geq 2$. In this section we assume $n = 2$, but we use a notation that allows the analysis to be generalised. Independently and without communication, each player sees the two labels and chooses one of them. Each receives a payoff of 1 if their choices *match* – that is, if both choose the same label – and a payoff of zero otherwise. Every label is unique, and so no two interactions are identical. However, to allow a simple representation of similarity between games, we assume that every label has one of two features, A and B, and that in each game, one label has feature A and the other has feature B. As modellers, *we* 'know' that the label with feature A in one game is (in this respect) similar to the label with feature A in another game. But we do *not* assume that players can directly map these features from one game to another, or even that they are conscious that every game has an A and a B label. When a player compares two games, it is a matter of subjective judgement for her how far (what we call) the A label in one game is similar to (what we call) the A label in another.

For example, suppose that in each game, the two labels are alternative videos of the behaviour of two vehicles facing the Right Turn Problem at some crossroads. Every game

has a unique pair of videos. In any given game, the only difference between the two videos is which vehicle gives way to which. Since the players' common objective can be interpreted as that of coordinating on an assignment of priority to one of the two vehicles, this example can be interpreted as a stylised model of a real-world Right Turn Problem.[6] Suppose that in each game, there is one video in which the second arrival gives way to the first (feature A), and one in which the first arrival gives way to the second (feature B). However, the fact that every pair of videos has this pair of features is not made explicit to the players. Across games, other features of the videos – such as whether the first arrival is going straight ahead or turning right, the types of vehicle involved, the flow of traffic on the main road, the weather, and so on – may vary.

As a model of individual behaviour, we postulate the following *replication heuristic*. The heuristic has two 'settings'. In any period in which a given individual i is using the *default setting*, that individual chooses between labels in some way that is independent of her experience of previous games. Her choice might be random, or it might be influenced by properties of the relevant pair of labels. We simply assume that, in any randomly selected game, the *default probabilities* with which a randomly selected player chooses A and B are q_A and q_B, where $0 < q_A, q_B < 1$ and $q_A + q_B = 1$. In period 1, each individual acts on the default setting.

In each period $t > 1$, each individual i uses the default setting if and only if she failed to match with her co-player in period $t - 1$; otherwise the *similarity setting* is operative. In this case, she tries to replicate the previous match by choosing whichever label in the period t game she perceives to be more similar to the label that she chose in the previous game. We model this process by defining measures r_A, r_B of the *intrinsic replicability* of the two features, where $0 \leq r_A, r_B < 1$. Intrinsic replicability encompasses the concepts of salience and reliability we introduced in Section 1.

For a randomly selected player in a randomly selected game, the probability with which A (respectively B) is chosen, conditional on that player having chosen A (respectively B) in the previous round and having matched with her co-player, is denoted by s_A

[6] One feature of the real-world problem that has been abstracted from the model is the conflict of interest between real drivers about *whose* vehicle has priority. In the game, players are not identified with particular drivers. Thus, our model is a pure coordination game rather than a Battle of the Sexes game.

(respectively s_B); these are measures of *gross replicability*. We model the process of replication by:

$$s_A = r_A + (1 - r_A)q_A, \quad \text{and} \tag{1}$$

$$s_B = r_B + (1 - r_B)q_B. \tag{2}$$

This specification ensures that s_A and s_B are strictly positive and that default probabilities have some influence on choice even when a player is trying to replicate a previous match; and it imposes the natural restriction that a player is at least as likely to choose a given type of label when trying to replicate another label of that type as when acting on default probabilities.[7]

The replication heuristic can be interpreted as a particularly simple and cognitively undemanding form of similarity-based inductive learning, in the same spirit as Gilboa and Schmeidler's (1995) case-based decision theory. Since it responds only to the success or failure of the individual's own actions, it does not involve any theory of mind or strategic reasoning. It requires only a one-period memory, and does not keep track of the relative success of alternative actions.

For any period t, let π_t be the relative frequency with which, in the whole population, A is chosen in that period. We define a function f such that, for all t, $\pi_{t+1} = f(\pi_t)$. Each game played in period t must have one of three outcomes – a match on A, a match on B, or no match. These outcomes occur with the respective probabilities π_t^2, $(1 - \pi_t)^2$, and $2\pi_t(1 - \pi_t)$. It follows immediately from the specification of the replication heuristic that

$$\text{for all } t\text{: } f(\pi_t) = \pi_t^{\,2} s_A + 2\pi_t(1 - \pi_t)q_A + (1 - \pi_t)^2 (1 - s_B) \tag{3a}$$

$$= \pi_t^{\,2}[1 + s_A - s_B - 2q_A] + \pi_t[2q_A + 2s_B - 2] + [1 - s_B]. \tag{3b}$$

A stationary state or *equilibrium* of the model is defined by a relative frequency $\pi^* \in [0, 1]$ such that $f(\pi^*) = \pi^*$.

To investigate the properties of this equilibrium, notice that f is a quadratic function with $f(0) = 1 - s_B$ and $f(1) = s_A$; f is everywhere convex (respectively: concave) if $1 + s_A - s_B - 2q_A$ is positive (respectively: negative). Equivalently, f is everywhere convex (concave) if

[7] This formulation allows the model to be extended to cases where $n > 2$. The general model has label types $j = 1, ..., n$, default probabilities $q_j \in (0, 1)$ and intrinsic replicability measures $r_j \in [0, 1)$. Following a match on j, j is chosen with probability $r_j + (1 - r_j)q_j$, while each $k \neq j$ is chosen with probability $(1 - r_j)q_k$.

$(s_A - s_B) - (q_A - q_B)$ is positive (negative). Given these properties of f, the following result is an immediate implication of the assumption that s_A and s_B are strictly positive:

> *Result 1*: There is exactly one equilibrium π^*. This equilibrium satisfies $0 < \pi^* < 1$ and is globally stable.

Figure 1 illustrates equilibrium as the intersection of the graph of $y = f(\pi)$ with the line $y = \pi$. (In the case illustrated, $q_A = q_B = 0.5$, $r_A = 0.6$ and $r_B = 0.2$, implying $\pi^* = 0.59$.)

[Figure 1 near here]

It follows from (1), (2) and (3a) that, for all $\pi \in (0, 1)$, $f(\pi)$ is increasing in r_A and q_A and decreasing in r_B, implying the following comparative static properties of equilibrium:

> *Result 2*: Other things being equal, π^* increases as r_A and q_A increase, and as r_B decreases.

In other words, the replication heuristic tends to favour label types that have higher default probabilities and greater intrinsic replicability. The following result gives some feel for the trade-off between decreases in default probability and increases in replicability:

> *Result 3*: π^* is greater than (equal to, less than) 0.5 if and only if $s_A - s_B$ is greater than (equal to, less than) $q_B - q_A$.

Proofs of Results 3, 4 and 5 are presented in the Appendix.

Our next result concerns the effects of varying the intrinsic replicability of A and B *together* while maintaining equality between r_A and r_B. Because of the symmetry between A and B, there is no loss of generality in considering only the case in which $q_A \geq 0.5$:

> *Result* 4: Assume $r_A = r_B = r$ and $q_A \geq 0.5$. Then:
> (i) if $r = 0$, $\pi^* = q_A$;
> (ii) if $q_A = 0.5$, then $\pi^* = 0.5$ for all $r \in [0, 1)$;
> (iii) if $q_A > 0.5$, then as $r \to 1$, $\pi^* \to 1$;
> (iv) if $q_A > 0.5$, then $\pi^* > q_A$ for all $r \in (0, 1)$.

This result shows that the overall effect of the replication heuristic is to magnify dispersion in the distribution of default probabilities, and hence to increase the frequency of coordination relative to the default benchmark. Obviously (given the symmetry between the

two label types, and given the assumption that they have equal intrinsic replicability r), if A and B have equal default probabilities, then they are chosen with equal frequency in equilibrium. But part (iv) of Result 4 establishes that if A has strictly greater default probability and if $r > 0$, the frequency with which A is chosen is not only greater than 0.5; it is strictly greater than the default probability q_A. Intuitively, this is because replication is activated by matching. If players choose according to default probabilities, the ratio between *choices* of A and *choices* of B is q_A:q_B, but the ratio of *matches* on A to *matches* on B is q_A^2:q_B^2. If $q_A > 0.5$, the latter ratio is greater than the former, and so replication works disproportionally in favour of A. The higher the value of r, the greater the effect of this disproportion on the equilibrium; as r tends to one, the equilibrium frequency of A choices tends to one also.

Finally, we consider whether each individual benefits by using the replication heuristic, rather than by acting on default probabilities. Consider any individual i. For the purposes of this analysis, we use q_A, q_B, r_A and r_B to denote the default probabilities and intrinsic replicabilities that are relevant in determining i's behaviour; we continue to assume $0 < q_A, q_B < 1$, $q_A + q_B = 1$, and $0 \leq r_A, r_B < 1$. Let $\pi \in [0, 1]$ be the (constant) relative frequency with which A is played in the population, and hence (given the assumption that the population is large) the probability that A is chosen by i's co-player in each game. We make no assumptions about the determinants of π or about the relationship between π and q_A. In particular, we do *not* assume that i's co-players have the same default probabilities as i does, nor that they use the replication heuristic. We assume only that their behaviour has some consistent pattern, described by π. Is i's expected payoff per round greater if her decision rule is the replication heuristic ('rule *R*') than if it is the use of default probabilities ('rule *D*')?

In general, the answer to this question depends on q_A, r_A, r_B and π. (For example, if $\pi > 0.5$ and $r_A < r_B$, replication works in the 'wrong' direction; if this effect is sufficiently strong, i might get a higher expected payoff by using *D* rather than *R*.) But a sharp answer is possible for all cases in which A and B have the same intrinsic replicability:

> *Result 5*: Consider any individual i for whom the default probabilities and intrinsic replicabilities of A and B are q_A, q_B and r_A, r_B, where $r_A = r_B = r > 0$. Suppose that, in every period, i's co-player chooses A with some constant probability $\pi \in [0, 1]$.

Let $v(D, t)$ be i's expected payoff in period t, conditional on her using rule D in all periods. Let $v(R, t)$ be i's expected payoff in period t, conditional on her using rule R in all periods. Then $v(R, t) \geq v(D, t)$ for all t. If $\pi \neq 0.5$, $v(R, t) > v(D, t)$ for all $t > 1$.

At first sight, it might seem surprising that, given only the assumptions $r_A = r_B > 0$ and $\pi \neq 0.5$, rule R can be shown to be unambiguously superior to rule D. The key to the proof is that, when individual i uses rule R, the effect of replication is always to increase the probability with which she chooses the label type that her co-players choose more frequently. This is the case because the replication heuristic does not try to replicate whatever co-players do; it tries only to replicate *matches*. For example, consider the case $\pi = 0.6$, $q_A = 0.9$, $r = 0.3$. If i tried to replicate her co-players' behaviour in general, the probability with which she chose A would tend to fall from its default level. But because the replication heuristic is activated only by matches, this probability will tend to increase. Suppose, for example, that in period 1, i matches with her co-player. The posterior probability that this match was on A is $(0.9)(0.6)/ [(0.9)(0.6) + (0.1)(0.4)] = 0.931$. Thus, the probability that i chooses A in period 2 is $(0.931)[0.3 + (0.7)(0.9)] + (0.069)(0.7)(0.9) = 0.909$, which is greater than q_A.

To sum up, we have described a very simple heuristic based on the principle of choosing actions that are similar to actions that have proved successful in the immediate past. When this heuristic is used by populations of players of recurrently similar pure coordination games, there is a tendency for the emergence of conventions that are based on shared perceptions of similarity. Other things being equal, this process tends to favour those putative conventions that have higher default probabilities and higher intrinsic replicability.

A convention that is based on labelling similarities works by picking out specific labels in individual games, and so can be interpreted as a source of salience, in Schelling's sense of the term (or, to be strictly accurate, in the sense of Schelling's use of the term 'prominence'). Thus, the emergence of such conventions can also be understood as the emergence of conceptions of salience. We suggest that the conceptions that are favoured by this process have at least something in common with those features of labels that distinguish focal points in one-shot coordination games. In the theoretical and experimental literature, one recurring idea is that focal points are grounded in *primary salience* – that is, in individuals' predispositions to choose some labels rather than others, in the absence of any strategic or payoff-related reasons to do so (Lewis, 1969; Mehta et al., 1994a; Bardsley et al., 2010). This idea can be developed by using *level*-k or *cognitive hierarchy* theories, in which

pre-strategic dispositions are attributed to players who reason at 'level 0' (Stahl and Wilson, 1995; Camerer et al., 2004; Crawford et al., 2008). The default probabilities of our model capture something of the same idea. Another recurring idea is that focal points are distinguished by properties of uniqueness in their labelling. Thus, in games with a finite number of labels, labels that are perceived as 'odd-ones-out' tend to be chosen, even if they are not primarily salient (Schelling, 1960; Bacharach and Stahl, 2000; Casajus, 2001; Janssen, 2001; Bacharach, 2006; Bardsley et al., 2010). It is plausible to suppose that, in families of games with clearly-defined odd-ones-out, the odd-one-out type of label will generally have high intrinsic replicability. Of course, we do not mean to suggest that salient solutions to artificial coordination games (such as '12 noon' in Schelling's 'Name a time to meet' game) have emerged from recurrent play *of those specific games*. But we do suggest that, in repeatedly dealing with a variety of real-world coordination problems, people may have learned that it is generally in their interests to choose strategies whose labels have obvious properties of uniqueness.

We do not claim that our simple model captures all the important features of the process by which, in reality, people learn to play recurrently similar coordination games. Our aim has been to demonstrate the feasibility of modelling similarity-based learning about labels, and to show that such learning imparts a tendency for the emergence of conventions based on replicable similarity relationships. There are many ways in which the model might be developed to make it more realistic, but we conjecture that this tendency would remain. For example, our simple model assumes that, after failing to achieve a match in any period, a player reverts with probability 1 to her default strategy. One alternative possibility would be to assume that, with some positive probability, such a player continues to try to replicate the most recent previous match. Another possibility would be to assume that, with some positive probability, a player who has failed to match in the previous period tries to replicate the choice that her co-player made in that period.[8] Either of these revisions would reduce the extent to which equilibrium was influenced by default probabilities, but they would not displace the crucial mechanism of the simple model, namely players' attempts to replicate

[8] Since the behaviour of a previous co-player provides evidence about the likely behaviour of the current co-player, each of these revisions can be interpreted as assuming that players are more rational than in the simple model.

one another's choices. Whenever such a mechanism is at work, there will be a tendency for emergent conventions to be based on those labelling features that are most replicable.[9]

4. Some experimental evidence

In this section, we discuss some evidence of learning behaviour in an experimental implementation of recurrently similar pure coordination games. We must emphasise that this experiment was not designed to test hypotheses deriving from the model presented in Section 2. To the contrary, the experiment was exploratory. Its aim was to investigate learning in recurrently similar coordination games, but no specific learning heuristics were chosen in advance to be tested. Our model was developed in the process of trying to understand the data generated by the experiment. The experiment is reported in another paper (Alberti et al., 2010). In this paper, we merely summarise its main features and describe some of the results that informed our modelling work.

The experiment involved 118 student subjects, randomly and anonymously assigned to pairs; pairings were maintained for the duration of the experiment. Fixed pairings were used to simplify the subjects' learning problem and to allow conventions to emerge relatively quickly. The experiment used forty pure coordination games. Each pair of subjects played all of these games, but not in the same order. Each game was defined by a set of four labels or *images*. Players were instructed to try to match with their co-players, and were paid in proportion to the number of matches they achieved.[10] After each game, players were told which images their co-players had chosen, thus allowing opportunities for them to learn from one another's behaviour. Games were presented in 'blocks' of five similar games; each pair of subjects played all the games in one block before moving on to another block. There were four blocks of *culture-laden* games and four of *abstract* games. Each pair of subjects either played all the culture-laden games before playing the abstract ones, or vice versa; which type of game was played first was counterbalanced.

---

[9] In other words, we conjecture that for a wide class of learning models, analogues of Results 1, 2 and (in the sense that increases in intrinsic replicability and decreases in default probability offset one another) Result 3 will hold. Results 4 and 5 depend on the more specific assumption that players try to replicate successes.

[10] In each experimental session, the payment per match was calculated ex post by dividing a total pool of £10 per subject in proportion to matches achieved. This payment mechanism gives each subject an incentive to achieve as many matches as possible, while giving no incentive for subjects to form collusive agreements before entering the experiment.

In each game in each culture-laden block, each of the four images was an example of a different *style*; the same four styles appeared in each game. However, this feature of the design was deliberately not made explicit to subjects, who were told only that every game was made up of four 'pictures'. Figure 2 shows two games from a block of culture-laden games in which the images are fabric designs and styles represent particular periods in the history of fashion in western society. In the figure, images with the same style appear above one another. In the experiment, however, the positions of the images were randomised independently for each co-player, so that position could not be used as a coordinating device. In two blocks the images were fabric designs; in the other two they were paintings. In the latter case, each style corresponded with a specific painter; within a game, the subject matter of the four paintings was (as far as possible) similar. Each block used a distinct set of four styles (fashion periods or painters). The images for all culture-laden games are available from the authors.

[Figure 2 near here]

Figure 3 shows two games from a block of abstract games. In these games, there were no pre-determined styles. Although the twenty abstract games were presented to subjects in blocks of five, all games were constructed on the same principles. Each image is a chequered pattern of coloured squares, using three distinct colours in a common pattern. In any given game, two *fixed colours* and their respective positions are held constant across all four images. The third *variable colour* is different in each of the four images. This design feature induces a general resemblance between the images in any given game. Subject to the constraints we have described (and the additional constraint that no two images in a given game should be identical) colours were selected at random from a pre-determined set of forty-eight colours, which had been constructed so that no two colours were difficult to distinguish from one another; the location of the variable colour was also selected at random. The images for all abstract games are available from the authors.

[Figure 3 near here]

Before playing the twenty abstract (respectively: culture-laden) games, each subject completed a questionnaire in which, in turn, she was shown eight sets of four abstract (culture-laden) images. For each set, she was asked to record which of the four images she 'liked most' and which she 'thought the person whom she was paired with liked most'. In fact (although this was not revealed to the subject at the time), these were the sets of images

that would appear in the first and last games that the subject played in each of the four blocks of abstract (culture-laden) games. Clearly, this questionnaire might have made subjects more conscious than they would otherwise have been of 'liking' as a rule for coordination. Thus if (as in fact was the case) there was a general tendency for subjects to choose the labels they most liked, that should not be used as evidence that liking is an important ingredient of salience in one-shot games.[11] But the aim of the experiment was to investigate how players' behaviour *evolved* over a series of recurrently similar games. For this purpose, the source of players' initial ideas about salience is immaterial. In fact, there was very little correlation between the likings of different subjects, indicating that in this experiment, liking would not be a successful coordination rule.[12]

For any given game, we define the *matching frequency* as the relative frequency with which *co-players* chose the same image. Following Mehta et al. (1994a), we define the *coordination index* for a game as the probability that two distinct players, *selected at random from the whole population of players*, chose the same image. This latter statistic uses individuals' actual choices, but (in contrast to the matching frequency) does not take account of who was paired with whom. As a benchmark, notice that (given that we are dealing with a four-label game) if all players choose at random, then the expected values of both the matching frequency and the coordination index are 0.25. The extent to which the coordination index exceeds 0.25 is a measure of positive correlation between choices *among subjects in general*. The extent to which the matching frequency exceeds the coordination index measures any additional *pair-specific* correlation.

Table 1 presents some summary statistics about matching frequencies and coordination indices for different types of game and at different stages in the experiment. We use the term 'round' to refer to the order in which games were played. Thus, for each set of twenty games of the same type (culture-laden or abstract), 'rounds 1–10' refers to the first ten games of that type faced by any pair of subjects, while 'rounds 11–20' refers to the remaining games of that type. Because the order in which games were faced by different pairs was randomised, comparisons between data for 'rounds 1–10' and 'rounds 11–20' pick up effects

[11] But probably it is. Bardsley et al. (2010) find a strong tendency for players of pure coordination games to use 'liking' (or 'favouriteness') as a means of coordination.

[12] For responses to questions about which image the subject liked most, the average coordination index (defined in the next paragraph) was 0.289 for abstract games and 0.287 for culture-laden games. Coordination indices for questions about co-players' likings were even closer to the random-response benchmark of 0.25.

of experience. The table also reports tests of whether matching frequencies are significantly greater than coordination indices, and whether coordination indices are significantly greater than 0.25.[13]

[Table 1 near here]

Notice that coordination indices, although consistently and significantly greater than the 0.25 benchmark, are typically quite close to 0.25, even in the later stages of the experiment. In contrast, matching frequencies are markedly and significantly greater than coordination indices. The increase in the matching frequency for abstract games between rounds 1–10 (0.361) and rounds 11–20 (0.423) suggests that, by using their experience of playing recurrently similar games, co-players were able to increase their success in matching. It is perhaps not surprising that a similar effect was not observed for culture-laden games, for which each five-game block had its own characteristics and styles.

It seems clear that some kind of similarity-based learning occurred. Since co-players were matched at random, the only credible explanation of the excess of matching frequencies over coordination indices is that subjects learned something from their co-players' behaviour, that this learning facilitated matching, and that different pairs followed different learning paths.[14] In a general sense, any learning that facilitates matching in recurrently similar games *must* exploit similarity relations between games: if players did not perceive similarities between the games they played, they would have no way of using what they learned in one game to guide their decisions in another. But the fact that learning was predominantly pair-specific provides a further clue about what was being learned. During the course of the experiment, the only feedback that each player received was information about her co-player's choices. So this information (and this information alone) must have been the input to pair-specific learning. That is, players must have adapted their own choices in response to their co-players' earlier choices. Since the overall effect of this adaptation was to increase

[13] The first test was carried out by using repeated random reassignments of subjects to pairs. For each such pairing, we calculated the average matching frequency (MF) implied by subjects' *actual* choices, thus generating a probability distribution of MF consistent with the null hypothesis of no pair-specific correlation. The second test used repeated simulation of random choices to generate a probability distribution of the coordination index consistent with the null hypothesis of random choice.

[14] In principle, an excess of matching frequencies over coordination indices could be an artefact of the formula for calculating coordination indices, which pools games played in different rounds. If, over the course of the experiment and independently of her co-player's behaviour, each subject became more skilled at playing the game, coordination indices would exceed matching frequencies. But the absence of any marked upward trend in coordination indices counts strongly against this explanation.

matching, and since different pairs followed different learning paths, the obvious inference is that in some way, players adapted their decision rules to favour choices that were similar to their co-players' previous choices.

Further evidence about this adaptation process is provided by the data on subjects' 'likings', as reported in the questionnaire. When playing their first culture-laden or first abstract game, subjects were very likely to choose the label that they most liked (the proportions were 0.686 for culture-laden games and 0.593 for abstract). The frequency with which most-liked images were chosen tended to decline over the twenty relevant games, but remained well above 0.25 throughout; and there was a spike at the start of each new block of games.[15] This pattern suggests that subjects used their own likings as a default rule, but that as they played more games of a given type, they adapted their own decision rules to favour choices similar to those that had been made by their co-players. The fact that the default rule was liking, rather than some other concept of salience, might perhaps be attributed to the questionnaire. But this rule was clearly not responsible for subjects' success in coordination. (Recall that there was very little correlation between subjects' likings, and that coordination indices were consistently low.) That success must have been due to pair-specific learning of *other* rules.

To make more direct inferences about the learning mechanisms that subjects used, we focus on some specific similarity relationships between games. For each block of culture-laden games, the four styles provide a pre-defined set of similarity relationships between games (analogous with features A and B in our model). We investigate how (if at all) subjects used these relationships. We do not assume that these are the only concepts of similarity that players can use; the significance of styles is merely that they allow us to identify *one* set of well-defined similarity relationships. We can then investigate whether there was any tendency for these particular similarity relationships to be used by players to increase the probability of coordination.

We now explain our method of analysis. Take any block of five similar culture-laden games and any given subject i. We define periods $t = 1, ..., 5$ to specify the order in which these games were played by i and her co-player. Take any of the four styles that is defined

[15] For more details, see Alberti et al. (2010). Which image a subject most liked was a better predictor of her choices than which image she thought her co-player most liked.

for these games, and denote this by *s*. For each period $t = 2, ..., 5$ we know whether *i* and/or her co-player chose style *s* in period $t - 1$, and whether *i* chose style *j* in period *t*. If *i* were using a similarity-based learning rule, and if she recognised the similarity relationship described by style *s*, we should expect that, other things being equal, she would be more likely to choose *s* in *t* if her co-player had chosen *s* in $t - 1$ than otherwise. If attempts at replication were activated only by immediately previous matches (as in the simple model presented in Section 2), this tendency would be restricted to cases in which *both* players had chosen *s* in $t - 1$.

Aggregating across all culture-laden blocks, all relevant styles *s*, all subjects *i* and all periods $t = 2, ...5$, there are $4 \times 4 \times 118 \times 4 = 7552$ 'observations'. Each observation can be classified by whether, in $t - 1$, *s* was chosen (i) by both *i* and her co-player, or (ii) only by *i*, or (iii) only by *i*'s co-player, or (iv) by neither. It can also be classified by whether, in *t*, *s* was or was not chosen by *i*. This cross-tabulation is reported in Table 2. For each of the cases (i) to (iv), the first column reports the total number of observations. The second column reports the absolute frequency of choices of style *s*. The third column reports the *unweighted* relative frequency of such choices, defined as the ratio of the entry in the second column to the entry in the first. The fourth column reports the *weighted* relative frequency of choices of style *s*, arrived at by calculating the relative frequency of such choices for each style separately and taking the mean. The weighted measure controls for the potential bias that, even in the absence of learning, matching will tend to be more common on styles that are more frequently chosen.

[Table 2 near here]

As one would expect, there is a tendency for subjects to replicate *their own* previous style choices, even when those previous choices had not lead to matches. The '*i* only' row of the table shows that, following periods in which two co-players failed to match, the weighted relative frequency with which subjects repeated their own previous style choices was 0.308, compared with the random-choice benchmark of 0.250. This may indicate that the default choices of styles by given subjects were positively correlated across games. (For example, a subject's default rule might be to choose according to her likings, and she might like some styles more than others.) Alternatively, it might indicate that some subjects were consciously trying to replicate their own previous choices as a means of facilitating coordination.

For our purposes, it is more important to see that subjects tended to replicate *their co-players*' style choices. As noted above, subjects who had failed to match in one period repeated their own previous style choices with a weighted relative frequency of 0.308. But for subjects who had matched, the corresponding frequency was 0.449. Another relevant comparison is between the 'neither' and 'co-player only' entries. When subjects had failed to match, styles which neither player had chosen in the previous round were chosen with a weighted relative frequency of 0.184 (per style: there were two such styles), while the style that the co-player had chosen in the previous round was chosen with a weighted relative frequency of 0.292.

To test for the significance of these effects, and to investigate whether some styles were more replicable than others, we carried out a random-effects probit regression analysis for each style. For each style *s*, the dependent variable was an individual's propensity to choose style *s* in period $t = 2, ..., 4$; the independent variables were a dummy to represent whether *i* had chosen *s* in $t - 1$ (*own choice*), and a dummy to represent whether *i*'s co-player had chosen *s* in $t - 1$ (*other's choice*).

The regression analysis revealed marked differences between styles. We will use the term 'Style 1.2' to refer to style 2 in block 1, and so on.[16] The *own choice* coefficient was significantly positive at the 5 per cent level in six cases – styles 2.2, 2.3, 2.4, 3.4, 4.3 and 4.4. In four of these cases (styles 2.2, 2.3, 2.4, 3.4) it was significant at the 1 per cent level. The *other's choice* coefficient was significantly positive at the 5 per cent level in seven cases – styles 2.1, 2.2, 2.3, 2.4, 3.3, 3.4 and 4.3. In six of these cases (all except style 3.3) it was significant at the 1 per cent level. The high levels of significance found for some styles, and the extent of overlap between the sets of styles that were significant in the two regressions, suggest that what we have found is *not* random variation. We conclude that there was a systematic tendency for subjects to replicate their co-players' style choices, but that this effect was highly style-specific: some styles were much more likely to be replicated than others. We tried adding an interaction term to the regressions to pick up the specific effect of a previous match, but this was statistically significant at the 5 per cent level in only one of the sixteen regressions.

[16] The games shown in Figure 2 are from block 2; from left to right, the four images in each game exhibit styles 2.1, 2.2, 2.3 and 2.4.

We now turn to the abstract games, which do not have pre-determined styles. We suggest that *one* salient way of distinguishing between the four images in a game is in terms of the variable colour. Thus, for example, in abstract game A2 (shown in the top row of Figure 3), the variable colours (from left to right) might be perceived as 'pale turquoise', 'purple', 'green' and 'dark blue'. Suppose that, in one round of the experiment, a pair of players face this game and match on the image on the left. In the next round, they face abstract game A3 (shown in the bottom row of Figure 3). If a player tries to replicate the previous match, which image in game A3 is most similar to the image chosen in game A2? One possible answer is that since the image chosen in game A2 was the palest of the four, the palest image in game A3 (presumably the third from the left) should be chosen.

By virtue of the way in which the experiment was computerised, each of the forty-eight colours is described by a unique triple of parameters (x_R, x_G, x_B) where $x_j \in [0, 1]$ is the intensity of colour j, and where j is one of the three primary colours for light (R = red, G = green, B = blue). These parameters allow us, as analysts, to define colour-based similarity rules in an objective way. For the purposes of our analysis, we define eight different *similarity rules*. One rule is 'choose the most red', which we operationalise as 'maximise $x_R/(x_R + x_G + x_B)$'. (Because colour mixes are defined in terms of light rather than pigment, a high value of $(x_R + x_G + x_B)$ represents a 'pale' or 'unsaturated' colour, that is, a colour close to white.) The rules 'choose the most green' and 'choose the most blue' are defined analogously. Similarly, the rule 'choose the least red' is operationalised as 'maximise $(1 - x_R)/[(1 - x_R) + (1 - x_G) + (1 - x_B)]$'. The rules 'choose the least green' and 'choose the least blue' are defined analogously. Colours at the 'least blue' extreme appear yellow; 'least green' colours appear purple, and 'least red' colours appear turquoise. Our final two rules use the unsaturated/saturated dimension. The rules 'choose the most pale' and 'choose the least pale' are operationalised as 'maximise $x_R + x_G + x_B$' and 'minimise $x_R + x_G + x_B$' respectively. For consistency with our discussion of culture-laden games, we will sometimes refer to 'most red', 'least pale' and so on as 'styles'; thus an image that is uniquely picked out by the rule 'choose the most red' has the style 'most red'.

Typically, each of these rules identifies a unique image in each game. (In approximately five per cent of cases, the maximisation or minimisation criterion of a similarity rule is satisfied by two or more images.[17]) Since there are eight similarity rules but

[17] For example, in game A3, the first and fourth images jointly satisfy the 'most blue' criterion.

only four images in each game, a given image is often picked by more than one rule. (For example, in game A2, image 3 is 'least pale', 'most green', and 'least blue'.) When this is the case, two co-players who use different similarity rules to replicate a given previous match are likely to fail to coordinate. It is thus particularly interesting to ask whether some colour-based similarity rules are more salient than others (in the sense discussed in Section 1).

Table 3 has the same structure as Table 2. Since each subject faced twenty abstract games of the same basic type, periods are defined by $t = 1, ..., 20$. The 17 022 'observations' are restricted to cases in which the relevant similarity rule identified a unique image both in period $t - 1$ and in period t, with $t \geq 2$. Notice that this implies that, if players choose at random, each style s will be chosen with probability 0.25 in all cases that are included in the analysis. The data show the same qualitative patterns as were found for culture-laden games: subjects tended to replicate both their own style choices and those of their co-players.

As in the case of culture-laden games, we carried out a random-effects probit regression analysis for each style. The interaction term between *own choice* and *other's choice* was significant at the 5 per cent level in only one regression and so was dropped. In the resulting equations, the *own choice* coefficient was significantly positive at the 5 per cent level in two cases – 'most blue' and 'least blue'. In one of these cases ('most blue') it was significant at the 1 per cent level. The *other's choice* coefficient was significantly positive at the 5 per cent level in five cases – 'most red', 'most green', 'most blue', 'least green' and 'least blue'. In two of these cases ('most blue' and 'least green') it was significant at the 1 per cent level. Again there seems to be a systematic but style-specific tendency for subjects to replicate their co-players' style choices.

Overall, the results of the experiment support the hypothesis that, when playing recurrently similar coordination games, individuals use similarity relations between the labels of successive games as a means of facilitating coordination. The crucial mechanism is that of players trying to choose strategies that are similar to those previously chosen by their co-players. The experimental results also suggest that some similarity relations are more likely to be used, or to be used successfully, than others. In these respects, the results support the general modelling strategy we have proposed.

However, we must acknowledge that our regression results do not support the hypothesis that players specifically try to replicate *previous matches*. Given that the repeating of previously successful rules or actions is the core mechanism of models of

reinforcement learning, this negative result is perhaps surprising. We suggest that it could be connected with the fact that the experiment used fixed pairs of players, rather than randomly reassigning subjects to pairs between games. This set-up may prompt subjects to try to anticipate their co-players' choices, rather than passively responding to successes. We must also report that, aggregating across all subjects, we were unable to find any systematic trends in the relative frequencies with which different styles were chosen over the five-game sequences of culture-laden games or over the twenty-game sequence of abstract games.[18]

5. Conclusion

We have shown how a simple form of experiential learning can lead to the emergence of conventions that are defined in terms of similarity relations between the labels that individuals use to describe games to themselves. We have also shown how the evolutionary selection of conventions can be influenced by the comparative replicability properties of different similarity relations. We do not claim that our stripped-down representation of experiential learning can explain all the ingredients of 'salience', understood as the defining characteristic of focal points in one-shot coordination games. But we suggest that it provides some clues about how conceptions of salience might begin to emerge, and hence about the possibility of a common source for those properties of salience that we can all recognise but which conventional game theory seems unable to represent.

Some readers of previous versions of this paper have suggested that, in our treatment of salience, we have merely shunted a well-known problem from one track to another, or kicked an unwanted can further down a road. We strongly disagree. As we have explained, most evolutionary game theorists have worked on the assumption – explicit or more usually implicit – that an analysis of conventions need take no account of labelling. They may recognise the truth of Schelling's hypothesis that properties of labelling are sometimes important for equilibrium selection in one-shot games, but they do not see that as relevant for evolutionary analysis. Why not? The answer is surely that, in the standard tool-box of evolutionary game theory, there is no plausible mechanism that could explain how labels affect the kinds of experiential learning that the theory is intended to represent. We have

[18] For example, the coordination indices for abstract games played in rounds 1–10 and 11–20 were 0.282 and 0.281 respectively (see Table 1). If subjects' choices had been converging on a limited number of styles, there would have been an upward trend in the coordination index.

proposed just such a mechanism. The two main components of this mechanism – default choice probabilities and the relative replicability of different labelling features – can be investigated empirically. What is more, they can be investigated in non-strategic settings.[19] So these concepts are not just re-descriptions of the fact that some strategies in a game are chosen more frequently than others: they are elements of a potential explanation of empirical phenomena. This explanation may turn out to be correct or incorrect; but unless salience is to be treated only as a mysterious footnote to game theory, potential explanations of its effects must be proposed and tested.

[19] Default choices can be investigated in experiments in which subjects are asked to 'just pick' one of a set of labels, with no incentive to pick any label rather than any other (Mehta et al., 1994a; Bardsley et al., 2010). Similarity judgements are an established topic of psychological research (e.g.Tversky, 1977); one investigative strategy would be to show subjects two sets of labels, pick out one label in one set and then ask subjects to report which label in the second set is most similar to it.

Table 1: Matching frequencies and coordination indices

	average matching frequency (MF) and coordination index (CI) for:			
	culture-laden games		abstract games	
	MF	CI	MF	CI
all games	0.369***	0.280***	0.392***	0.284***
culture-laden block 1 (fabrics)	0.356*	0.316***		
culture-laden block 2 (fabrics)	0.407***	0.263***		
culture-laden block 3 (paintings)	0.346***	0.280***		
culture-laden block 4 (paintings)	0.366***	0.263***		
games played in rounds 1–10	0.367**	0.282***	0.361*	0.298***
games played in rounds 11–20	0.370**	0.281***	0.423***	0.274***

Asterisks in the MF columns report bootstrap tests of the hypothesis that MF is higher than CI. Asterisks in the CI columns report bootstrap tests of the hypothesis that CI is greater than 0.25 (10, 5 and 1 per cent significance are shown by *, ** and *** respectively).

Table 2: Replication of styles between periods: culture-laden games

	number of observations	style *s* chosen by player *i* in period *t*:		
		absolute frequency	unweighted relative frequency	weighted relative frequency
in *t*–1, *s* chosen by:				
i and co-player	682	356	0.522	0.449
i only	1206	378	0.313	0.308
co-player only	1206	362	0.300	0.292
neither	4458	792	0.178	0.184
total	7552	1888	—	—

Table 3: Replication of styles between periods: abstract games

	number of observations	style *s* chosen by player *i* in period *t*:		
		absolute frequency	unweighted relative frequency	weighted relative frequency
in *t*–1, *s* chosen by:				
i and co-player	1888	696	0.369	0.354
i only	2697	796	0. 295	0.293
co-player only	2697	769	0. 285	0.282
neither	9740	2144	0. 220	0.224
total	17022	4405	—	—

Figure 1: Equilibrium

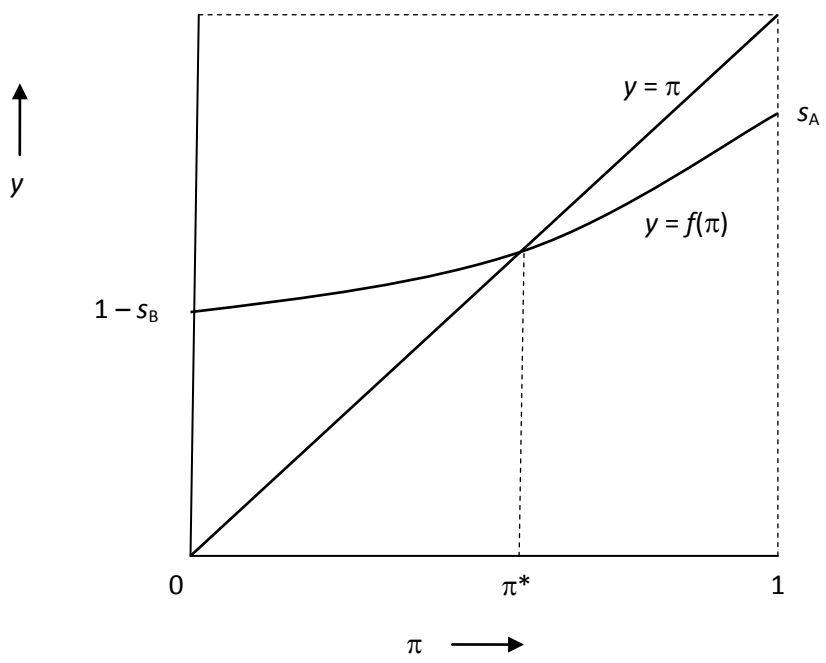
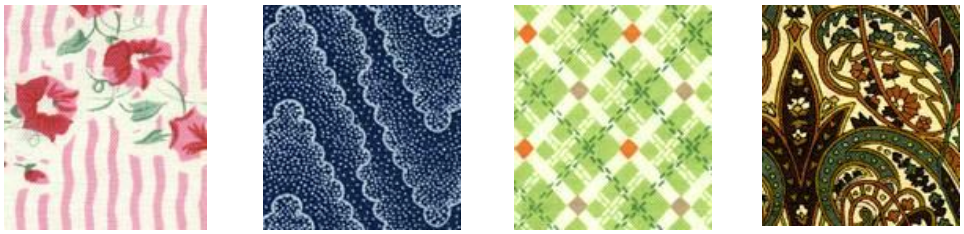


Figure 2: Two culture-laden games

game C6



game C7

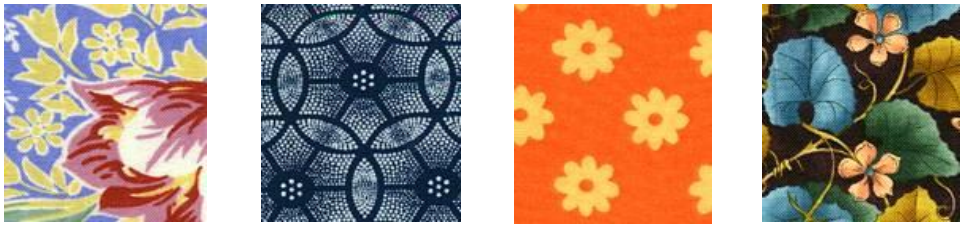
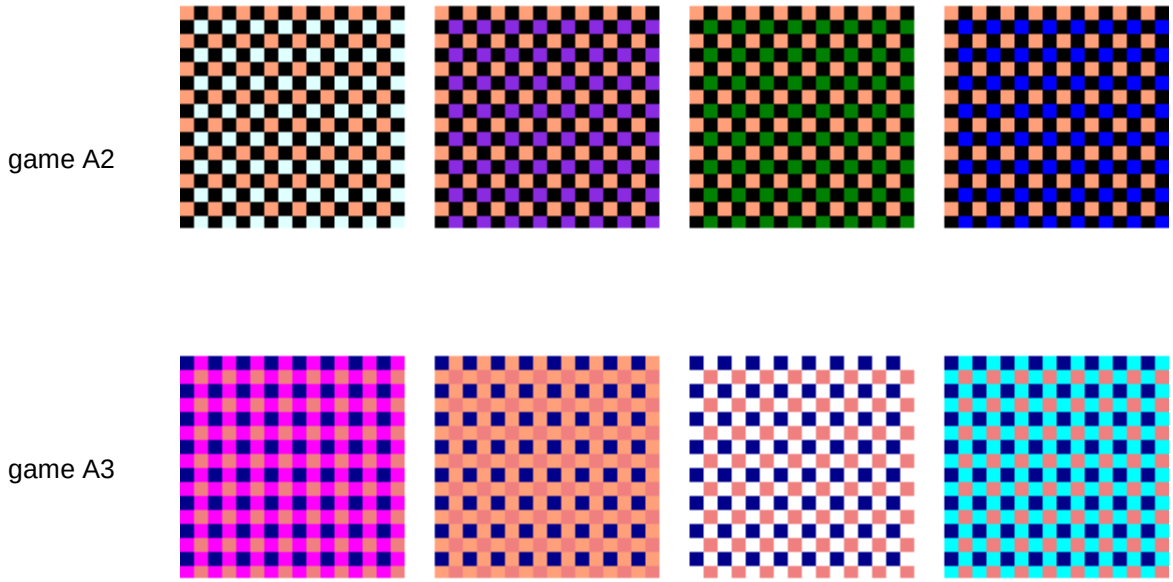


Figure 3: Two abstract games



The variable colours (x_R, x_G, x_B) in the four A2 images are, from left to right: (0.88, 1.0, 1.0), (0.54, 0.17, 0.89), (0.0, 0.5, 0.0) and (0.0, 0.0, 1.0). The variable colours in the A3 images are (1.0, 0.0, 1.0), (1.0, 0.63, 0.48), (1.0, 1.0, 1.0) and (0, 1.0, 1.0).

References

Alberti, Federica, Shaun Hargreaves Heap and Robert Sugden (2010). The emergence of salience: an experimental investigation. CBESS Discussion Paper 10-17, Centre for Behavioural and Experimental Social Science, University of East Anglia.

Bacharach, Michael (2006). *Beyond Individual Choice: Teams and Frames in Game Theory* (Natalie Gold and Robert Sugden, eds). Princeton, NJ: Princeton University Press.

Bacharach, Michael, and Michele Bernasconi (1997). The variable frame theory of focal points: an experimental study. *Games and Economic Behavior* 19(1): 1–45.

Bacharach, Michael, and Dale O. Stahl (2000). Variable-frame level-*n* theory. *Games and Economic Behavior* 33(2): 220–46.

Bardsley, Nicholas, Judith Mehta, Chris Starmer, and Robert Sugden (2010). Explaining focal points: cognitive hierarchy theory *versus* team reasoning. *Economic Journal* 120 (March): 40–79.

Binmore, Ken and Larry Samuelson (2006). The evolution of focal points. *Games and Economic Behavior* 55: 21–42.

Camerer, Colin F., Teck Ho and Kuan Chong (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics* 119(3): 861–98.

Casajus, André. (2001). *Focal Points in Framed Games: Breaking the Symmetry*. Berlin: Springer-Verlag.

Crawford, Vincent P., Uri Gneezy, and Yuval Rottenstreich (2008). The power of focal points is limited: even minute payoff asymmetry may yield large coordination failures. *American Economic Review* 98 (4): 1443–1458.

Cubitt, Robin P, and Robert Sugden (2003). Common knowledge, salience and convention: a reconstruction of David Lewis's game theory. *Economics and Philosophy* 19(2): 175–210.

Gauthier, David (1975). Coordination. *Dialogue* 14: 195–221.

Gilboa, Itzhak, and David Schmeidler (1995). Case-based decision theory. *Quarterly Journal of Economics* 110 (3): 605–639.

Goyal, Sanjeev and Maarten C.W. Janssen (1996). Can we rationally learn to coordinate? *Theory and Decision* 40: 29-49.

Harsanyi, John C. and Reinhard Selten (1988). *A General Theory of Equilibrium Selection in Games*. Cambridge, Mass.: MIT Press.

Hayek, Friedrich A. (1973). *Law, Legislation and Liberty. Volume 1: Rules and Order*. Chicago: University of Chicago Press.

Hume, David (1739–40/ 1987). *A Treatise of Human Nature*. Oxford: Clarendon Press.

Isoni, Andrea, Anders Poulsen, Robert Sugden and Kei Tsutsui (2011). Focal points in tacit bargaining games. CBESS Working Paper 11-13, University of East Anglia.

Janssen, Maarten C.W. (2001). Rationalising focal points. *Theory and Decision* 50(2): 119–48.

Lewis, David K. 1969. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Mehta, Judith, Chris Starmer, and Robert Sugden (1994a). The nature of salience: an experimental investigation of pure coordination games. *American Economic Review* 84(3): 658–73.

Mehta, Judith, Chris Starmer, and Robert Sugden (1994b). Focal points in pure coordination games: an experimental investigation. *Theory and Decision* 36(2): 163–85.

Rankin, Frederick W., John B. Van Huyck and Raymond C. Battalio (2000). Strategic similarity and emergent conventions: evidence from similar stag hunt games. *Games and Economic Behavior* 32: 315–337.

Roth, Alvin E. and Ido Erev (1995). Learning in extensive-form games: experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior* 8: 164–212.

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

Schelling, Thomas C. (1978). *Micromotives and Macrobehavior*. New York: Norton.

Schlicht, Ekkehart (1988). *On Custom in the Economy*. Oxford: Oxford University Press.

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

Stahl, Dale O. and Wilson, Paul W. (1995). On players' models of other players. *Games and Economic Behavior* 10(1): 218–54.

Sugden, Robert (1995). A theory of focal points. *Economic Journal* 105 (May): 533–550.

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

Sugden, Robert (2011). Salience, inductive reasoning and the emergence of conventions. *Journal of Economic Behavior and Organization*, 79: 35–47.

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

Tversky, Amos (1977). Features of similarity. *Psychological Review* 84: 327–352.

Young, H. Peyton (1998). *Individual Strategy and Social Structure*. Princeton, N.J.: Princeton University Press.

Appendix 1: Proofs of results

Results 1 and 2 are proved in the main text.

Proof of Result 3

It is evident from Figure 1 that π^* is greater than (equal to, less than) 0.5 if and only if $f(0.5)$ is greater than (respectively: equal to, less than) 0.5. From (3a):

$$f(0.5) > (=, <)\ 0.5 \text{ iff } 0.25s_A + 0.5q_A + 0.25(1 - s_B) > (=, <)\ 0.5.$$

Since $q_A + q_B = 1$, this can be rewritten as:

$$f(0.5) > (=, <)\ 0.5 \text{ iff } 0.25(s_A - s_B) > (=, <)\ 0.25(q_A + q_B) - 0.5q_A$$

or:

$$f(0.5) > (=, <)\ 0.5 \text{ iff } s_A - s_B > (=, <)\ q_B - q_A.$$

Proof of Result 4

Assume $r_A = r_B = r$ and $q_A \geq 0.5$.

To prove part (i), assume $r = 0$. Then $s_A = q_A$ and $s_B = 1 - q_A$. Thus, using (3a), $f(\pi) = q_A$ for all $\pi \in (0, 1)$. Since equilibrium is defined by $f(\pi^*) = \pi^*$, $\pi^* = q_A$.

To prove part (ii), assume $q_A = 0.5$. Then $s_A - s_B = q_B - q_A = 0$. Thus, by Result 3, $\pi^* = 0.5$.

To prove part (iii), assume $q_A > 0.5$. As $r \to 1$, $s_A, s_B \to 1$. Thus, using (3a):

$$\text{for all } \pi \in (0, 1): \text{ as } r \to 1,\ f(\pi) - \pi \to \pi^2 + 2\pi(1 - \pi)q_A - \pi. \qquad \text{(A1)}$$

But

$$\pi^2 + 2\pi(1 - \pi)q_A - \pi = \pi(2q_A - 1)(1 - \pi), \qquad \text{(A2)}$$

which (given $q_A > 0.5$) is positive for all $\pi \in (0, 1)$. Thus, in the limit as $r \to 1$, $f(\pi) > \pi$ at all $\pi < 1$, implying $\pi^* \to 1$.

To prove part (iv), it is sufficient to prove that if $r > 0$,then $q_A > 0.5$ implies $f(q_A) - q_A > 0$. From (3b):

$$f(q_A) - q_A = q_A^2[1 + s_A - s_B - 2q_A] + q_A[2q_A + 2s_B - 3] + [1 - s_B]. \qquad \text{(A4)}$$

Using (1) and (2) and rearranging:

$$f(q_A) - q_A = q_A\, r\, (2q_A - 1)(1 - q_A), \qquad \text{(A5)}$$

which is strictly positive if $q_A > 0.5$ and $r > 0$.

Proof of Result 5

Notice that if $\pi = 0.5$, then $v(D, t) = v(R, t) = 0.5$ for all t. Notice also that, because the two decision rules prescribe the same behaviour in period 1, $v(D, 1) = v(R, 1)$. Thus, given the symmetry between A and B, it is sufficient for a proof of Result 5 to show that if $r > 0$ and $\pi > 0.5$, then $v(R, t) > v(D, t)$ for all $t > 1$.

Assume $r > 0$ and $\pi > 0.5$. Consider any period $t > 1$. First suppose that *i did not* match with her co-player in period $t - 1$. Then, irrespective of whether she is using rule *D* or rule *R*, she chooses A in period t with probability q_A, and so her expected payoff in t is the same for both rules.

Now suppose instead that *i did* match with her co-player in $t - 1$. If she is using rule *D*, she chooses A in t with probability q_A. But suppose she is using rule *R*. For periods $T = 1, ..., t$, let ρ_T denote the probability with which *i* chose (or chooses) A in T. First, we show (*Lemma* 1) that $\rho_{t-1} \geq q_A$ implies $\rho_t > q_A$.

Assume $\rho_{t-1} \geq q_A$. Conditional on *i* having matched in $t - 1$, the probability that this match was on A is given by:

$$m_{t-1} \equiv \rho_{t-1}\, \pi / [\rho_{t-1}\, \pi + (1 - \rho_{t-1})(1 - \pi)]. \qquad \text{(A6)}$$

Thus:

$$m_{t-1} > q_A \;\Leftrightarrow\; \rho_{t-1}\, \pi / [\rho_{t-1}\, \pi + (1 - \rho_{t-1})(1 - \pi)] > q_A. \qquad \text{(A7)}$$

The right-hand side of (A6) is increasing in ρ_{t-1}. Thus, since $\rho_{t-1} \geq q_A$,

$$q_A\, \pi / [q_A\, \pi + (1 - q_A)(1 - \pi)] > q_A \;\Rightarrow\; m_{t-1} > q_A\,. \qquad \text{(A8)}$$

But $\pi > 0.5$ implies that the antecedent of (A8) holds, proving that $m_{t-1} > q_A$.

From the assumption that *i* uses rule *R*:

$$\rho_t = m_{t-1}\, [r + (1 - r)q_A] + (1 - m_{t-1})(1 - r)q_A$$

$$= q_A + r(m_{t-1} - q_A). \quad \text{(A9)}$$

Since $m_{t-1} > q_A$ and $r > 0$, (A9) implies $\rho_t > q_A$, proving Lemma 1.

Maintaining the assumptions that $t > 1$, that i matched in $t - 1$, and that i is using rule R, we define k as the number of successive periods, immediately prior to t, in which i matched. (In other words: i matched in periods $t - 1$, $t - 2$, ..., $t - k$; and *either* $t - k = 1$ *or* i failed to match in $t - k - 1$.) Suppose $k = 1$, i.e. i failed to match in $t - 2$. Then $\rho_{t-1} = q_A$, and so $\rho_t > q_A$ by Lemma 1. Now suppose $k = 2$. Then $\rho_{t-2} = q_A$, and so $\rho_{t-1} > q_A$ by Lemma 1. Applying Lemma 1 again, $\rho_t > q_A$. This argument can be repeated for $k = 3, 4, ...$, establishing $\rho_t > q_A$ for all k.

If $\pi > 0.5$, i's expected payoff in any given period is higher, the higher the probability with which she chooses A. If i uses rule D, she chooses A with probability q_A in all periods. We have established that if $\pi > 0.5$ and if i uses rule R, the probability with which she chooses A in any period $t > 1$ is strictly greater than q_A if she matched in $t - 1$, and is equal to q_A otherwise. Since q_A, $r \in (0, 0)$, the probability of matching when using rule R is non-zero in all periods. Thus $\pi > 0.5$ implies $v(R, t) > v(D, t)$ for all $t > 1$, completing the proof.