Are There Lewis Conventions?*

Francesco Guala
University of Exeter (UK) and San Raffaele University (Italy)
f.guala@ex.ac.uk

Abstract

David Lewis famously proposed to model conventions as solutions to coordination games, where equilibrium selection is driven by precedence, or the history of play. A characteristic feature of Lewis Conventions is that they are intrinsically non-normative. Some philosophers have argued that for this reason they miss a crucial aspect of our folk notion of convention. It is doubtful however that Lewis was merely analysing a folk concept. I illustrate how his theory can (and must) be assessed using empirical data, and argue that it does indeed miss some important aspects of real-world conventions. I conclude that whether Lewis Conventions exist or not depends on how closely they approximate real-world behaviour, and whether we have any alternative theory that does a better job at explaining the phenomena.

You are sitting in front of a computer screen. Using your mouse, you can choose one of two coloured buttons labelled, from left to right, “Red” and “Blue”. You know that two other players are facing the same decision. If you all choose the same colour, you will earn 10 experimental tokens each, which will be converted later into real money. Unfortunately you have to make your decision simultaneously, without the possibility of communicating with the other group members. You also know that you will play this game ten times with the same partners, and will receive feedback after each round. What will you choose?

It seems that in the first round you cannot do better than choosing at random. But in fact, unbeknown to you, your body is already helping you out. Like most people, when the screen appeared in front of you, you probably fixated your sight on the button placed on the left-hand side of the screen. You then shifted your sight to the right-hand button, returned to the left, and repeated this process several times. Eventually, there is a higher probability that you will choose the object upon which you fixated first (see Rangel 2007).

So with a bit of luck all the players in your group will choose Red and earn 10 tokens already in the first round. But even if this does not happen, two players out of three will necessarily choose the same colour. This will send a message to the third player. Using a

* Research for this paper was made possible by the ESRC grant RES-000-22-1591 and the Computable and Experimental Economics Laboratory of the University of Trento. Previous versions were presented at the University of Amsterdam, a seminar of the British Society for Philosophy of Science, and the annual meetings of the Italian Society for Analytical Philosophy (SIFA) and the Italian Society for the History of Economic Thought (STOREP). I’m grateful to members of these audiences and in particular to Ivan Moscati, Mario Gilli, and Philippe Mongin for their feedback. The usual disclaimers apply.

simple majority rule, she will infer that choosing that colour is the most likely coordination strategy in the next round. Following this reasoning, your group should be able to coordinate in just a few rounds, and from then on rather effortlessly make money by simply repeating the choice made in the previous round.

At this point a *convention* has emerged. David Lewis (1969) first proposed to model conventions as solutions to repeated coordination problems of this kind. We can represent a simple coordination game using a standard two-by-two matrix (Table 1). You are the row player and for simplicity the other two members of the group are jointly represented as column. This game has two Nash equilibria: Red/Red and Blue/Blue. Standard game theory assigns an equal chance for Red and Blue to become coordination points in repeated play. Even worse, it is unable to predict that all players will keep playing the convention, once they have coordinated. But as a matter of fact, when this game is played in the laboratory two-thirds of the participants play Red in the first round, which is then twice as likely as Blue to evolve into a convention.[1] And of course the overwhelming majority continue to coordinate successfully after this has been done at least once.

	Red	Blue
Red	10, 10	0, 0
Blue	0, 0	10, 10

Table 1: A simple coordination task

Lewis borrowed the idea of modelling conventions as coordination games from Thomas Schelling (1962). Schelling had argued that in solving coordination problems we are often helped by apparently irrelevant factors that make one of the available strategies *salient*. Consider for example the "Ten Numbers" game: you must choose one among the following numbers:

0, 1, 2, 3, 4, 5, 6, 7, 8, 9

Your partner is sitting in a separate room and is facing the same problem. It is a one-shot game: if you both choose the same number, you will gain $10 each, otherwise nought. In a game like this, the probability of converging on the same option by playing randomly is very small. Yet, a surprisingly high number of people coordinate successfully by choosing zero. There are a number of factors that contribute to make zero salient: it is the first number in the list, and it is notoriously a peculiar number too. It is also the first one on the left, and as we have seen it is more likely to be chosen for purely physiological reasons.

A salient strategy constitutes a *focal point* that facilitates coordination when purely rational considerations are insufficient to pin down the best strategy. Focal points may be determined by cultural, cognitive, or even biological factors. Lewis argued that in the case of conventions salience is determined by *precedence*. Why do Britons drive on the

[1] In a sample of 141 experimental subjects, 93 chose Red in the first round, and 94 were playing Red after eight rounds.

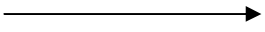
left? Forget the traffic code or the police: except a few fools, nobody drives on the left for fear of sanctions; we do it because we do not want to crash into one another. If everybody else were to swap from left to right, we would do the same, regardless of the law. In Britain we drive on the left because every driver has been doing it in recent history, and we expect them to continue to do so in the future.

Lewis' account was remarkable for a number of reasons. It pioneered the application of game theoretic tools in the field of social ontology. It introduced the concept of common knowledge, and highlighted the importance of repeated play – an insight that has recently been vindicated by the development of evolutionary game theory. Finally, it exposed the limitations of "pure" rational choice theory for the analysis of collective behaviour. If we want to understand how institutions emerge from individual interaction, we must study the ways in which cognitive, cultural, and biological biases constrain our behaviour, make it more predictable, and hence reduce the enormous complexity of social interaction. To constantly engage in the calculations of a perfectly rational player would be too time consuming, perhaps impossible for cognitively limited creatures as we are. Thus the a priori project of modelling perfectly rational players can only take us so far in the study of social behaviour. The study of conventions is inevitably an *empirical*, as well as a theoretical task.

Conventions and norms

In one important respect Lewis' theory sits firmly in the rational choice tradition. Our main motivation to follow a convention is strictly selfish: we drive on the left because we want to avoid accidents; we say "cat" rather than "tac" because we want to be understood by our interlocutors; we wear black at funerals because we want to communicate our grief. Lewis' approach then leads naturally to a neat separation between social norms and conventions. A social norm always comes with an intrinsic "ought", and is usually backed up by a system of sanctions. The sanctions are meant to change the payoffs of the game: for example, to change a mixed-motives game (like a prisoner's dilemma) into a coordination game (Figure 1).[2]

	Left	Right
Left	2, 2	0, 3
Right	3, 0	1, 1



	Left	Right
Left	2, 2	0, 0
Right	0, 0	1, 1

Figure 1: Transforming a Prisoner's Dilemma game into a Coordination game.

The transformation of (3, 0) and (0, 3) into (0, 0) may take place in different ways. If the payoffs represent utility values, as it is often the case in standard game theory, then the reduction of the "free-riding" payoffs (Right-Left and Left-Right) may be due to a feeling of guilt or shame: the other player had trusted my cooperation and I have let her down, for example. But in many societies there are external mechanisms that reduce our payoffs both at the psychological and at the material level: a verbal reproach or ostracism from

[2] For a seminal game-theoretic account of social norms along these lines, see Ullmann-Margalit (1977).

business are examples of how normative pressure helps attaining socially superior equilibria in the game of life.

Roughly, then, a social norm exists when every individual (1) prefers to conform to the norm provided that (almost) everybody else does the same; (2) it is common knowledge that one ought to conform; and (3) this normative expectation is backed up by sanctions.[3] Lewis somewhat misleadingly claims that "conventions are a species of norms". But Lewis Conventions are *not* norms in the sense specified by conditions (1)-(3). Rather, conventions are supported by *extrinsic* normative considerations: one follows a convention because (a) it is individually rational to do so, and (b) deviance from conventions is usually sanctioned by *other* independent social norms. A convention does not, *per se*, imply a commitment to conformity to the same strategy. While satisfying (1) and (2), condition (3) does not apply. The sanctions that support a convention are not tailored to supporting *that* particular strategy, but derive from considerations of a much more general kind.

The key paragraph from Lewis (1969) is worth quoting in full:

> we do presume, other things being equal, that one ought to do what answers to his own preferences. And we presume, other things being equal, that one ought to do what answers to others' preferences, especially when they may reasonably expect one to do so. For any action conforming to any convention, then, we would recognize these two (probable and presumptive) reasons why it ought to be done. We would not, so far as I can tell, recognize any similarly general reasons why it ought not to be done. This is what I mean by calling conventions a species of norms. (p. 98)

Notice that Lewis' expectations are "plain" expectations, to use Margaret Gilbert's (1989) expression, i.e. non-normative expectations about what others *will* do (as rational individuals), rather than what they *ought* to do. Lewis does not explain why one should answer to others' preferences in such situations. He only says that not doing so is likely to cause feelings of disapproval, and even to trigger sanctions (pp. 99-100). So we should imagine that breaking conventions would amount to violating some independent norm, like "do not harm others unless there is a good reason to do so" (Gilbert 1989, p. 354). Although "'convention' itself, on my analysis, is not a normative term", says Lewis, "there are certain *probable consequences* implied by the fact that an action would conform to a convention [...] which are presumptive reasons, according to our common opinion, why that action ought to be done" (1969, p. 97, emphasis added).

Lewis' analysis is controversial. Margaret Gilbert (1989, 2008) has argued forcefully that conventions, norms, and related social institutions (customs, traditions, rules) must be analysed in terms of more primitive notions of group action and collective intention. In particular, conventions result from a "quasi-agreement" among members of a group to pursue a certain line of action that will attain a specific collective goal. Such quasi-

[3] Although Lewis does not analyze social norms in depth, "Ludovician" theories of norms can be found e.g. in Pettit (1990) and Bicchieri (2006).

agreements need not be formulated explicitly, and often derive from the mere observation that people do pursue a certain line of action that serves the goals of the relevant group. Collective intentions result in a *joint commitment* that cannot be unilaterally breached by an individual group member. This is why, according to Gilbert, we usually feel the need to excuse and justify a breach of convention in front of other group members. One of Gilbert's complaints is that "conventions in Lewis' sense do not seem apt to give rise to the 'ought' judgments typically associated with conventions as ordinarily conceived" (1989, p. 354).

Theories of group action are sophisticated and are becoming increasingly influential, but this is not the place to examine them in detail.[4] Lewis' approach conflicts with these accounts in a number of ways. Gilbert even disputes that coordination games provide a good starting point for a philosophical analysis of convention. In what follows I will bracket such issues and focus on the main disagreement concerning normativity.[5] Even if coordination games did not provide necessary conditions for social conventions, they would still model a number of situations that we commonly associate with conventions. These "Lewis Conventions" – a technical term from now on – are the focus of this paper. But *are there* any Lewis Conventions, after all?

Analysis and intuitions

It is not clear how this question should be tackled. Lewis has been commonly read as providing an analysis of the vernacular notion of convention. Accordingly, critics like Gilbert have focused on counterexamples that exploit inconsistencies between his theory and the everyday conceptual apparatus associated with convention. Luckily, she claims, "we can tell much that we need to know about concepts by telling science fiction tales and such" (Gilbert 1989, p. 10). Here's one such tale:

> People in a certain community regularly take tea at four in the afternoon. Though this is population common knowledge no one affects a particular positive attitude towards the practice, beyond generally conforming to it. In particular, it is not regarded as mandatory in any way. When Sally suggests to Charles that he come for tea at five, Charles may be a little surprised but has no sense of impropriety. If this is the way things are I suggest that we would not say that they have a convention that four o'clock is the time to have tea. (Gilbert 1989, p. 350)

Let us take Gilbert's suggestion seriously: would *we* say that there is a convention to have tea at four, or not? It is hard to say. Linguistic practices do not constrain the usage of terms like "convention" enough for there being a definite answer to this question.[6]

[4] See Gold and Sugden (2007) for a critical analysis and overview.
[5] Notice that other philosophers do not associate the notion of collective intentions strictly with the idea of joint commitment. Searle (1990) and Bratman (1993), for example, have proposed non-normative theories of collective intentionality. Sugden's (2000) and Bacharach's (2006) theories of "team reasoning" also reject Gilbert's notion of commitment and restrict normativity to the "ought" of logical inference.
[6] Concerns of this kind are not new and are not peculiar to the philosophy of social science. They have emerged first in epistemology (see e.g. Stich 1990) and ethics (Horowitz 1998). For general surveys of

Philosophers' tales often stretch our intuitive capacities to the breaking point (as in the quoted paragraph) where clear intuitions are hard to come by.

Of course this conceptual gymnastic is far from uninteresting. In telling us what a convention *really* is Gilbert constructs a complex conceptual structure that is bound to be partly revisionary of the way in which we use our language. The logical positivists pointed out a long time ago that philosophical analysis can (and perhaps ought to) have a critical as well as a descriptive function.[7] But then agreement with our linguistic practice or with our intuitions in highly fictional scenarios cannot be the ultimate test of validity for philosophical reconstructions of folk concepts.[8]

Indeed, it may be more important to come up with a new, coherent concept of convention than trying to mirror a muddled discourse. In a recent contribution to social ontology Raimo Tuomela (2002) for instance declares to be interested in analyzing the "common-sense framework of [collective] agency". This framework is presented as the carrier of a great amount of useful information about social reality, and as an important testing device for philosophical constructs. However, he admits that ultimately the common-sense framework is likely to be incoherent. Only by revising it we can construct a coherent system that may help future social scientists:

> the resulting account [of social reality] does not really compete with what social scientists are doing as it rather is meant in part to critically analyze the presuppositions of current scientific research and [...] to provide a new conceptual system for theory-building (Tuomela 2002, p. 7)

Scientific theories, I take, must then be tested in the usual way. Ontological investigation can play a heuristic role, but is eventually appraised on the basis of the science it has produced. The ultimate validation must be empirical, rather than conceptual, in character.

Analytical empiricism

There are reasons to believe that Lewis himself would not disagree. Lewis (1969) says repeatedly that he is providing an analysis of convention. What he does *not* claim, however, is that he is primarily interested in providing an analysis of our folk notion of convention. While expressing the *hope* that it captures the vernacular concept of convention, Lewis is adamant that agreement with such a concept is neither the only nor the most important criterion for the appraisal of his theory:

---

recent work in so-called "experimental philosophy" see also Knobe (2006), Alexander and Weinberg (2006).

[7] See e.g. Reichenbach (1938, pp. 3-6). On revisionary metaphysics in general, see Carrara and Varzi (2001).

[8] Here I am consciously conflating "pure" concept analysis with the analysis of everyday usage of concepts, as revealed by linguistic practice. Philosophers disagree about the nature of concepts, and this is not the place to resolve such disagreements. Notice that language analysis can (and should, in my view) be an empirical activity, so in a sense I am here simply advocating the replacement of one kind of empirical data (how people use the concept/term "convention") with another kind of data (are there such things as conventions in Lewis' sense).

> I hope it is an analysis of our common, established concept of convention […]. But perhaps it is not, for perhaps not all of us do share any one clear general concept of convention. At least, insofar as I had a concept of convention before I thought twice, this is either it or its legitimate heir. And what *I* call convention is an important phenomenon under any name (Lewis 1969, p. 3; see also p. 46 for a reiteration of this point).

The analysis of folk theories of course plays an important role in Lewis' philosophy in general. One of Lewis' lasting contributions consists precisely in clarifying a method of philosophical analysis (the "Carnap-Ramsey-Lewis" method) that is applicable to a wide range of folk theories – from psychology, to mathematics, colours and even holes. So readers may have been misled into thinking that the project pursued in *Convention* is analogous to the analyses that Lewis provides elsewhere. But this is doubtful, and the best way of seeing this is by trying to place the theory of conventions in the context of Lewis' method.

In "Psychophysical and Theoretical Identifications" Lewis (1972) gives a detailed account of the method of analysis of folk theories. The analysis proceeds in four steps. First, collect all the "platitudes" of the folk theory in question. In the case of psychology, for example, the platitudes are going to be everyday principles like "if people want an object, believe that the object is within their reach, and no counteracting reason intervenes, then they try to grab that object", and other trivialities of this sort.

Second, form the conjunction of these platitudes.[9] This conjunction will include both problematic, "Theoretical" terms (mental states, for example), and unproblematic "Old" terms referring to familiar objects and phenomena (facial expressions, linguistic utterances, etc.). Following Carnap, Lewis proposes that the meanings of the T-terms be defined by their functional role in the folk theory – their relations with one another and with the O-terms of the theory. (For this reason, Lewis calls the conjunction of platitudes "the postulate of the term-introducing theory".)

All the T-terms can now be replaced with variables, and these variables can be quantified over to obtain claims of the form: "There are X, Y, Z,… that stand in such-and-such relations among themselves and with the O-terms". This quantified version of the conjunction of platitudes is the "Ramsey-sentence" of the folk theory. By "Ramseyfing" we explicate the role of problematic T-terms, simply by showing what their job is in the overall economy of the folk theory. Although the Carnap-Ramsey-Lewis approach has been widely debated, these three preliminary steps are meant to capture the core activities that most philosophers associate with the method of conceptual analysis. "Collecting the platitudes" actually gives a false appearance of simplicity to what is typically a difficult, controversial task. Counterexamples and "fiction tales" play a prominent role in deciding which platitudes are to be included among the postulates, and the definition of the folk

[9] I am simplifying here (and elsewhere) for ease of presentation. See Lewis (1970, 1972) for the full account.

theory is achieved by a difficult balancing act between general principles and intuitions about specific cases.[10]

Frank Jackson (1998) has argued that conceptual analysis is instrumental to the goals of "serious metaphysics". The Ramseyfication of a folk theory, in other words, should not be pursued as an end in itself. Serious metaphysics must bring order and simplicity in the heterogeneous list of what there is – the list of entities and properties that figure in our folk theories. The fourth step in the Carnap-Ramsey-Lewis method in fact is concerned with *reduction*, whereby problematic T-terms are shown to be co-referential with the less problematic terms of a base theory. In many cases – like the mind-body problem that concerns Lewis (1970, 1972) – the reduction is potential rather than actual. We do not know yet what the T-terms of folk psychology refer to, although presumably future neuroscience will let us know. In the meantime we can still say something general about the denotation of the folk concepts, by explicating the causal roles that brain states will have to account for, in order to attain the reduction of mental states.

Successful completion of this four-step process hinges crucially on the strength of the analysandum, that is, on the correctness of the folk theory in question. In the case of psychology we seem to have a decisive advantage, for we have direct access to the folk theory in question. Lewis goes as far as to saying that the principles of folk psychology are common knowledge (albeit of the tacit kind) and therefore only require to be made explicit for all members of the folk to recognize their validity. This advantage can be used to pull out a simple trick. Consider that the Ramsey-sentence implies the theory: if X, Y, Z exist then the theory is true. The latter implication (or "Carnap-sentence": $R_T \supset T$) is analytic in Lewis' view. Lewis introduces a "modified Carnap sentence" to ensure uniqueness: on pain of indeterminacy of reference, the theory must refer to one set of entities only. And here comes the trick: if the modified Carnap-sentence is analytic, then obviously *either* the T-terms do not refer, *or* our platitudes about them are true. But if the folk theory has been analyzed properly, then the platitudes *are* true (they are platitudes after all!). So the T-terms do refer (although we may not know exactly *what* they refer to).

Lewis uses this trick explicitly in his work on the mind-body problem. The T-terms are names of mental states, and the O-terms name sensory stimuli, motor responses, and the like. Once our folk-psychology has been Ramseyfied, we know what sort of job the entities that will replace mental states in our future base theory must do – even though we do not know exactly what these entities are. Lewis follows more or less the same strategy in his work on colours and the foundations of mathematics,[11] but the case of conventions is more complicated. Unlike folk psychology, vernacular social ontology is hardly common knowledge among the folk. On the contrary, if social psychologists are right we should expect it to be deeply mistaken on a number of issues, and in a systematic way

[10] Jackson (1998) offers a book-length exposition and defense of the analytical method that owes much to Lewis' work.
[11] According to Nolan (2005).

too.[12] With such a problematic analysandum, the Carnap-Ramsey-Lewis project cannot even take off.

And in fact there is an important disanalogy between Lewis' approach in *Convention* and his method of analysis of folk theories. The key T-term ("Lewis Convention", as we shall call it) is defined by Lewis using a *scientific model* rather than a set of folk platitudes. The model is partly borrowed from the theory of games, and is partly of Lewis' own invention. There is no doubt that Lewis believes that many platitudes can be captured by his theory – and yet the platitudes do not constitute the theory itself.

There are, to sum up, two possible interpretations of Lewis' project. On one reading, he is indeed attempting an analysis of our folk notion of convention. He is concerned with the first three steps of the Carnap-Ramsey-Lewis method, in other words. And yet, consider the O-terms: far from relying on unproblematic notions, Lewis analyzes convention using sophisticated concepts such as utility maximization, Nash equilibrium, and common knowledge. Once this has been done, the Ramseyfication of Lewis' theory of conventions will not deliver the trick. According to the analytic (modified) Carnap-sentence, either Lewis Conventions do not exist, or the theory is true. But since the theory is not just a conjunction of platitudes, it may well be false. Lewis Conventions may not exist after all.

On another reading, Lewis is proposing a *scientific theory* that may (or may not) provide the base for the reduction of "folk" conventions. He is concerned with the last step (reduction) of the Carnap-Ramsey-Lewis method, in other words. Of course we cannot guarantee that a scientific theory is able to capture all the features of folk conventions. We may have to be eliminativist regarding at least some of the latter. But this may not matter if, as Lewis says, "what *I* call convention is an important phenomenon under any name" (1969, p. 3).

Under the first reading, Lewis can be criticized for doing an imperfect analysis of the folk notion of convention. His theory does not fit our core intuitions. This is Gilbert's interpretation, and should be dismissed in my view.[13] According to the second interpretation, the question of the correctness of Lewis' theory is a *scientific* one. Consider an analogy with physics: the reduction of thermodynamics to molecular physics is predicated on the fact that the latter gets most things right, at its own level of analysis. The discovery that the motion of particles can do (almost all) the job of temperature is exciting precisely because the laws governing this motion are secure on experimental grounds. Similarly, the reduction of mental states to brain states will occur only when the principles of neurophysiology will be properly understood and validated. Has this prerequisite been satisfied in the case of conventions? If Lewis' theory were not confirmed by empirical data, then it would not even be a contender for metaphysical reduction. If the theory did not describe the phenomena adequately at its own level of

[12] See for example Rothbart and Taylor (1992).

[13] Gilbert (2008) has argued recently that Lewis' theory can be interpreted *both* as an attempt to analyze a folk concept, *and* as a descriptive account of a real-world phenomenon. Although this is surely a move in the right direction, I still believe it to be false for the reasons outlined in this section. Concept analysis is at best a secondary goal for Lewis (1969).

analysis, then the issue of whether we have good intuitions about, say, normativity, would not even arise. We wouldn't have to choose between a scientific and a folk theory, if the scientific theory was imperfect or even plainly false. That's why, according to this reading, Lewis' theory must be assessed in the laboratory, rather than in the philosopher's armchair.

Back to the lab

We have seen that coordination is achieved quite easily in small groups playing repeatedly the game in Table 1. But this does not mean that Lewis was right. Lewis Conventions involve a particular set of mechanisms that facilitate and support coordination, and the mere observation that coordination takes place sheds little light on the underlying mechanisms. Are experimental subjects driven by the motives highlighted by Lewis, or is there a more complicated story to be told? In particular, was Lewis right on normativity? Do instrumental rationality and external norms provide an exhaustive account of the "ought" of convention, or is there an intrinsic normative pressure to conform?

We can answer these questions by manipulating the incentives of the game. Suppose that after nine rounds of "normal" coordination play, the tenth and final round includes a surprise: instead of the incentive structure of Table 1, players will face the payoffs in Table 2. Whatever convention evolved in the early stages of the game (Red-Red or Blue-Blue), one player (we shall call her the "potential deviant") has an incentive to deviate from it. In Table 2 the potential deviant is the row player, and as usual the other two members of the group are jointly represented as column. The key feature is that by breaching the convention a deviant imposes a penalty on the other group members.

	Red	Blue
Red	200, 200	300, 0
Blue	300, 0	200, 200

Table 2: Incentive to deviate in the 10th round

Before the tenth round the potential deviant is informed about this change in the payoff structure, but the other two group members are not. She is told that they are not aware of this change, but that at the end of the game they will be fully informed about the payoff structure and the choice of the potential deviant. So before the tenth round the potential deviant can safely assume that the two other players will continue to follow the convention. As a potential deviant, your choice-situation is very simple: either *conform* (everybody earns 200) or *breach* the convention (you earn 300, they earn nothing).

We can now detect the effect of norms by observing whether potential deviants are willing to forego individual earnings and conform to the convention that evolved in earlier rounds. *The normativity of convention is the (normative) expectation that you ought to bear the possible costs of non-deviance, because I am planning my choices*

based on the expectation that you will conform. Normativity is manifested in the decision to "leave some money on the table" and privilege the group's earnings with respect to one's own private gain.[14]

Our tenth round can then be used as an "acid test" to detect the influence of normative forces that may have emerged during the early rounds of group play. As a matter of fact in the laboratory less than one-third (30%) of the potential deviants decide to breach the convention. This may sound remarkably low, but in fact is consistent with a large body of experimental data.[15] So why do people decide to conform? Notice that the incentive structure of the tenth round is similar to a sequential game in which the first mover does not have any other option except to put herself in the hands of the potential deviant. After the tenth round the game is over and the three players will never meet again. So a purely consequentialist, looking-forward agent will not be afraid of disrupting the convention that has emerged in the previous rounds.

If strategic considerations cannot play a role here, the remarkably high level of conformity with the convention can be explained by the existence of social norms that prescribe cooperation. A norm of *altruism* ("you ought to help the members of your group") for example may prescribe to conform to the established regularity. If it is common knowledge in the group that the norm applies to situations of this kind, the potential deviant may be willing to comply with the norm at the expense of some individualistic gain. Similarly, norms of fairness or equality will prescribe to conform to the behaviour of other group members because this is the way to achieve an equal distribution of the resources.[16]

We know that these norms are at work, because we can manipulate them. By changing the payoffs in the last round it is possible to trigger (or shield) different norms that dictate cooperation. Consider the payoffs in Table 3. In the tenth round now the potential deviant faces a straight choice between earning 300 at others' expenses, and letting them earn

[14] This use of monetary incentives is very common in experimental psychology and economics. In experiments with Prisoner's Dilemma or Ultimatum games, monetary incentives are used to observe the factors that prompt individuals to deviate from the predictions of standard rational choice theory. By "standard theory" I mean the theory of rational play based on Nash equilibrium, together with the assumption of self-interest that is commonly included in economic models. It is sometimes pointed out that game theory does *not*, strictly speaking, predict that individuals maximize their expected monetary payoffs. The theory says that individuals act so as to maximize their expected *utility*, and the latter does not have to be an increasing function of their monetary gains only. Utility, however, is not directly observable. By observing deviations from the prediction of the standard model we can try to reconstruct utility functions using behavioural evidence. This strategy is potentially fruitful, and has led to the creation of increasingly sophisticated models incorporating normative considerations of altruism, fairness, equality, and reciprocity. Philosophical and methodological discussion of these issues can be found in Bicchieri (2006), Guala (2006), and Woodward (forthcoming).

[15] See Guala and Mittone (2008) for a more detailed discussion of the experimental results, as well as footnote 16 below.

[16] The same behaviour can be captured by postulating other-regarding preferences. Although models of other-regarding preferences are popular among economists in virtue of their simplicity and tractability, they are known to have a number of defects. I will not pursue this distinction here, but a thorough discussion of the difference between theories of social norms and theories of other-regarding preferences can be found in Bicchieri (2006, Ch. 3).

200 at her own expense. Instead of a choosing between an individualistic and a cooperative outcome, as in Table 2, she now chooses among an altruistic and an individualistic outcome. Unsurprisingly, the proportion of deviants is much higher in this condition. 65% of players now choose a different colour and take the 300 tokens. This is to be expected if, as plausible, only a minority of altruists is willing to donate money at their own expense.

	Red	Blue
Red	0, 200	300, 0
Blue	300, 0	0, 200

Table 3: Altruistic vs. individualistic option

The design so far does not allow one to discriminate between "external" norms of cooperation and the "intrinsic" normative force of convention. When the potential deviant approaches the tenth round, she may be influenced by the history of play that has developed in her own group. Repeated team play over the first nine rounds may have generated an extra pressure to conform, over and above the considerations listed above. Perhaps, as suggested by Gilbert, this pressure is the consequence of a joint commitment associated with collective intentionality. In any case, the mere fact that some subjects are willing to conform to the convention and forego individual gains merely tells us that there is *some* norm at work, but does not indicate exactly *what kind* of normativity we are dealing with.

If the intrinsic normativity of convention emerges via repeated group play, we should be able to observe the net effect of external norms prescribing cooperation simply by eliminating group play. We can subtract the intrinsic force of convention, and leave only the effect of external norms. This is what happens in the one-shot game represented in Figure 2.

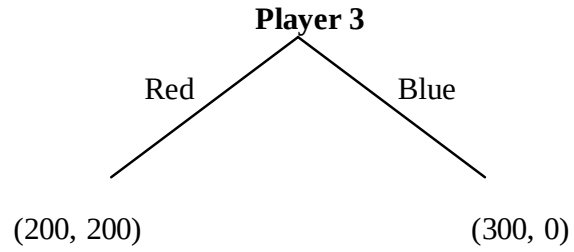


Figure 2: The one-shot game.

The decision tree represents the game as seen from the viewpoint of the potential deviant ("Player 3"). The first two players do not move – their colour is arbitrarily assigned by a computer. The potential deviant can observe the result, and then decide whether to play

the same colour or not. Notice that at this point she is facing exactly the decision situation of the tenth round of the repeated game, except that there is no history of group play, and thus no opportunity for the intrinsic normativity of convention to emerge. Whatever expectations are formed regarding the potential deviant's behaviour, they must arise from external social norms prescribing cooperation in situations of this kind.

When the one-shot sequential game is played in the laboratory, 68% of the experimental subjects decide to deviate, compared to 30% in the repeated game.[17] The mere fact of playing together for nine rounds is sufficient to enhance conventional behaviour. Conventions are not only sustained by external norms of cooperation, but also by an intrinsic normative pressure to conform to an established regularity.

Are there Lewis Conventions?

A Lewis Convention solves a coordination problem by acting as a focal point that guides our choices in future play. In Lewis' model each player follows the convention for two sets of reasons: to pursue her own selfish gain; and because external social norms dictate not to hurt others, ceteris paribus. Both reasons motivate the behaviour of real players facing simple choices in laboratory settings. But a third factor also influences real decision-makers. When a group of players build a history of joint action, they unintendedly create an additional pressure towards conformity that goes beyond the "ought" of individual rationality and the "ought" of external social norms. Whether this intrinsic normativity is to be explained by a joint commitment or some other mechanism is an important question that we do not know how to answer yet. More data must be collected to disentangle the complex causal processes underlying the dynamics of group play. For the time being, we can say that Lewis' model overlooks these processes and provides only a partial account of the ontology of conventions.

The experiments were designed to deliver a particularly powerful message. In real life, admittedly, we do not always interact anonymously with a group of strangers whom we are unlikely ever to meet again. But consider that our anodyne experimental settings are much *less likely* to create social pressure on the participants, than the sort of situations we face in everyday life. And yet, the intrinsic normativity of conventions can be observed even in these unfavourable conditions. We can only expect the pressure to *increase* when we play indefinitely repeated games with our family members, friends, and colleagues.

[17] This replicates what we already know from similar experiments. Charness and Rabin (2002) for instance have found remarkably similar results in a two-player sequential game where the first mover chooses between opting out and staying in the game. If she opts out, she will earn nothing and the first mover will earn 800 tokens; if she stays in, the second mover has a choice between taking all the money (0, 800) or sharing in equal parts (400, 400). In their sample, no first mover opts out, 56% of the second movers choose the "fair" outcome, and 44% choose the inequitable one. The importance of history is apparent also in Charness and Rabin's game. In another condition experimental subjects are offered a straight choice between the two allocations, (0, 800) and (400, 400). Technically, this is a mini-version of a so-called Dictator's game, where the other player (the equivalent of the "first mover", in the sequential game) is not allowed to make any decision whatsoever. In the Mini-dictator's game players opt in majority for the inequitable division (78%). So the mere fact that the first movers are allowed to do something and choose to stay in the sequential game is sufficient to shift almost 35% of the subjects towards the equitable outcome.

Thus "Lewis Convention", as a theoretical term, strictly speaking does not refer. To conclude that "folk" conventions do not exist would have a whiff of absurdity – surely we cannot question their existence, for we deal with conventions all the time. That's why some philosophers, contemplating the prospect that Lewis Conventions may not exist, suggest that they were a bad idea right from the start. Since Lewis did not capture the everyday notion of convention, surely we should let his theory rest in peace? This would be too hasty. If Lewis was not analyzing a folk theory, his theory should not be appraised with criteria that are appropriate to the analysis of folk theories. The relevant criteria are *scientific*, and Lewis' theory should be assessed in the light of these only. Intuitions do play a role in the test of social scientific theories, but they are not the evidence against which such theories are tested. They rather work as *heuristic* devices, suggesting mechanisms and hypothesis which must then be investigated empirically.

One final point is worth making: I have said that Lewis' theory is false, *strictly speaking*. But "strictly speaking" is too strict: *all* scientific theories are false to some extent, as far as we know. If literal truth was our criterion of appraisal, then no theoretical terms would refer, even in the most advanced sciences. There would be no quarks, electrons, atoms, chemical elements, molecules, cells, organisms, and so on and so forth. This seems to result in too much waste: the physical, chemical, and biological theories used to define these concepts are too important and successful to make referential success hostage to literal truth. Lewis (1970) calls the entities named by the T-terms of a theory the "realizers" of T. If there are no exact realizers of any important scientific theory, then we should only require that a theory is *nearly realized* by a set of entities, in order for its T-terms to refer. Or, at any rate, that it is *more nearly* realized than its rivals. Lewis' theory has some rivals, and some rival accounts (like Gilbert's)[18] seem to capture some details of the story that are overlooked by Lewis.

It is still early days, of course, and we should suspend judgment until more data have been gathered to test these alternatives. Even then, some difficult choices will lie ahead: to specify a metric of realization is a notoriously difficult problem in the philosophy of science. An adequate metric may have to combine different criteria on various dimensions, in the skilful and ingenious ways that scientists master. Although we are still far from cracking all these problems, it would be foolish to abandon the task. What conventions are is a scientific, empirical question, and we should invest our energies into answering it in a proper, scientific way. Lewis' theory gave us the conceptual framework and the methodological tools to pursue this project. It seems appropriate, then, to answer our question as follows:

> Are there Lewis Conventions?
> Probably not, but luckily we had Lewis' *Convention*.

References

[18] I use the term "account" to signal the fact that not all of them are as precisely and formally defined as Lewis' own theory.

Alexander, J. and Weinberg, J.M. (2006) “Analytic Epistemology and Experimental Philosophy”, *Philosophy Compass* 2: 56-80.

Bacharach, M. (2006) *Beyond Individual Choice: Teams and Frames in Game Theory*. Princeton: Princeton University Press.

Bicchieri, C. (2006) *The Grammar of Society*. New York: Cambridge University Press.

Bratman, M. (1993) “Shared Intention”, *Ethics* 104: 97-113.

Carrara, M. and Varzi, A.C. (2001) “Ontological Commitment and Reconstructivism”, *Erkenntnis* 55: 33-50.

Charness, G. and Rabin, M. (2002) “Understanding Social Preferences with Simple Tests,” *Quarterly Journal of Economics* 117: 817-69.

Gilbert, M. (1989) *On Social Facts.* London: Routledge.

Gilbert, M. (2008) “Social Convention Revisited”, *Topoi* 27: 5-16.

Gold, N. and Sugden, R. (2007) “Collective Intentions and Team Agency”, *Journal of Philosophy* 104: 109-137.

Guala, F. (2006) “Has Game Theory Been Refuted?” *Journal of Philosophy* 103: 239-63.

Guala, F. and Mittone, L. (2008) “An Experimental Study of Conventions and Norms”, CEEL Working Paper, University of Trento.

Horowitz, T. (1998) “Philosophical Intuitions and Psychological Theory”. In M. DePaul and W. Ramsey (eds.), *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*. Lanham, Md: Rowman and Littlefield.

Jackson, F. (1998) *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Oxford University Press.

Knobe, J. (2006) “Experimental Philosophy”, *Philosophy Compass* 2: 81-92.

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

Lewis, D.K. (1970) “How to Define Theoretical Terms”, *Journal of Philosophy* 67: 427-46. Reprinted in *Philosophical Papers, Vol. 1.* Oxford: Oxford University Press.

Lewis, D.K. (1972) “Psychophysical and Theoretical Identifications”, *Australasian Journal of Philosophy* 50: 249-58. Reprinted in *Papers on Metaphysics and Epistemology*. Cambridge: Cambridge University Press.

Nolan, D. (2005) *David Lewis*. Bucks: Acumen.

Pettit, P. (1990) “*Virtus Normativa*: Rational Choice Perspectives”, *Ethics* 100: 725-55.

Rangel, A. (2007) “The Role of Visual Attention in Simple Choices”, California Institute of Technology.

Reichenbach, H. (1938) *Experience and Prediction*. Chicago: University of Chicago Press.

Rothbart, M. and M. Taylor [1992] “Category Labels and Social Reality: Do We View Social Categories as Natural Kinds?”, in *Language, Interaction and Social Cognition*, edited by G.R. Semin and K. Fiedler. London: Sage.

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

Searle, J.R. (1990) “Collective Intentions and Actions”, in Cohen, P.R., Morgan, J. and Pollack, M.E. (eds.) *Intentions in Communication*. Cambridge, Mass.: MIT Press.

Stich, S. (1990) *The Fragmentation of Reason.* Cambridge, Mass., MIT Press.

Sugden, R. (2000) “Team Preferences”, *Economics and Philosophy* 16: 174-204.

Tuomela, R. (1995) *The Importance of Us*. Stanford: Stanford University Press.

Ullmann-Margalit, E. (1977) *The Emergence of Norms*. Oxford: Clarendon Press.

Woodward, J. (forthcoming) “Experimental Investigations of Social Preferences”, in H. Kincaid and D. Ross (eds.), *The Oxford Handbook of the Philosophy of Economics*. New York: Oxford University Press.