The Normativity of Lewis Conventions

Francesco Guala

University of Milan

(Forthcoming in *Synthese*)

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 nonnormative. 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.

Acknowledgments: 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 universities of Amsterdam, Cambridge, Exeter, and Trento; the Institut d'Histoire et de Philosophie des Sciences in Paris, a seminar of the British Society for Philosophy of Science, 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, Philippe Mongin, Giacomo Sillari and an anonymous referee for their detailed feedback. The usual disclaimers apply.

The Normativity of Lewis Conventions

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 decisions simultaneously, without the possibility to communicate 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. All this information is common knowledge among players. What will you choose?

It seems that in the first round you cannot do better than choose randomly. But in fact, unbeknown to you, your body is already helping you. 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 Armel et al. 2008).

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 it does not happen, at least two players out of three will necessarily choose the same colour. This will send a message to the player that chose differently. Using a simple majority rule, she will infer that switching 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 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 in pure strategies: Red/Red and Blue/Blue.

Standard game theory[1] 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 choosing 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. And of course the overwhelming majority continues to coordinate successfully after they have done it at least once. [2]

	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 (1960). 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 of 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 several factors that contribute to making zero salient: it is the

[1] By "standard" (as opposed to *evolutionary*) game theory, I mean the theory of strategic decision-making with fully rational agents that stemmed from von Neumann and Morgenstern's work in the 1940s and 50s.

[2] In a sample of 141 experimental subjects, 93 chose Red in the first round, and 94 were playing Red after eight rounds. Some details about the experimental design can be found in the Appendix.

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 this reason alone.

A salient strategy constitutes a *focal point* that facilitates coordination when purely rational considerations are insufficient to identify the best move. 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 Italians drive on the right? Forget the traffic code or the police: except a few fools, nobody drives on the right for fear of sanctions; we do it because we do not want to crash into one another. If everybody else were to swap from right to left, we would do the same, regardless of the law. Italians drive on the right because all drivers have been doing it in recent history, and they expect others to continue in the future.

To capture this intuition, Lewis defined convention as a behavioural regularity (R) in a recurrent situation such that (1) there is a history of conformity with R, (2) there are mutual expectations of conformity, (3) everyone prefers to conform with R, if (almost) everybody else does the same, and (4) everyone would prefer to conform with an alternative regularity R', if everybody else did the same. Crucially, these conditions must be common knowledge among the members of the community (for the full definition see Lewis 1969, p. 76).

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 interactions, 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 decision-making. 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.[3]

Conventions and norms

Our main motivation to follow a convention is self-interested: 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. Knowing these motivations, other individuals form expectations regarding our future moves. But these are non-normative (or "plain") expectations – they concern what other players *will*, rather than what they *ought* to do.

This approach thus seems to invite a neat separation between social norms and conventions. Cristina Bicchieri (2006, p. 38) for example claims that conventions never run counter to self-interested motives,[4] and only require plain expectations regarding others' behaviour. Social norms in contrast always come with normative expectations, and are usually backed up by sanctions (Bicchieri 2006, p. 11). The sanctions are meant to change the payoffs of the game: for example, to transform a mixed-motives game (like a prisoner's dilemma) into a coordination game (Figure 1).[5]

[3] The same point may apply to evolutionary models and simulations with boundedly rational players which, nevertheless, abstract from cognitive, cultural, and biological biases (see e.g. Skyrms 1996). Some biases, like framing effects, arguably can only be studied using experiments.

[4] Notice that rational choice theory does *not*, strictly speaking, assume that individuals maximize their material or 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. Unless otherwise stated, I will use the term "self-interest" to refer generically to motives directed towards the maximization of expected utility. "Selfish" instead will refer to behaviour aimed specifically at maximizing one's own material payoffs.

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

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

Roughly then a social norm exists when every individual prefers to conform to a behavioural rule or regularity R, provided that: (1) (almost) everybody else conforms; (2) there are plain expectations of conformity; (3) there are normative expectations that one ought to conform, and these normative expectations are sometimes backed up by sanctions (Bicchieri 2006, Chapter 1).[6] In contrast a convention in Lewis' sense does not, *per se*, imply a commitment to conform to the salient strategy. Lewis' expectations are "plain" expectations, to use Margaret Gilbert's (1989) expression, non-normative expectations about what others *will* do (as rational individuals), rather than what they *ought* to do.[7] While (1) and (2) are satisfied, condition (3) does not seem to apply.

This neat distinction however seems to clash with Lewis' own account. In a convoluted section, Lewis argues that "conventions may *be* a species of norms" (1969, p. 97

[6] See also Pettit (1990) for an account of norms along similar lines.

[7] There is still, of course, a prudential "ought" (what we believe that others should do, in order to be rational). So it would be more precise to say that "plain" expectations do not include any normative element apart from the normativity of instrumental rationality.

emphasis in original). His argument for normativity moves from two propositions (6 and 7, p. 98) that are implied by his analysis of conventions as solutions to coordination games. Every time I face a situation governed by an established convention, Lewis says,

> (6) I have reason to believe that my conforming would answer to my own preferences.
> (7) I have reason to believe that my conforming would answer to the preferences of most other members […] and that they have reason to expect me to conform.
> And (6) and (7), when true, are presumptive reasons why I ought to conform. For we do 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 a convention a species of norms (1969, p. 98).

Lewis goes on to say that failure to conform is likely to elicit negative reactions from the other players, and perhaps even sanctions. So Lewis Conventions look a lot like norms, all things considered.

Yet, an important difference remains. While social norms are *intrinsically* normative (you should not steal, even if it is in your interest to do so), conventions are not. The normative power of Lewis' conventions relies on two "external" sources: on the one hand, the normativity of conventions is the "ought" of instrumental rationality. On the other, it is the "ought" of norms that prescribe not to damage other individuals, other things being equal (or absent a countervailing reason to do so). One follows a convention because (a) it is individually rational to do so, and (b) deviance from conventions is prohibited by *other* independent moral principles or social norms.[8]

[8] An insightful and persuasive analysis of these pages can be found in Gilbert (1989, p. 354).

According to some philosophers, this kind of "extrinsic" normativity is too weak. Margaret Gilbert (1989, 2008) for example 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 collective goal. Such quasi-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.[9] Lewis' approach clashes with these accounts in a number of ways. Gilbert for example disputes that coordination games provide a good starting point for a philosophical analysis of convention. But even if coordination games did not provide necessary conditions for social conventions, [10] they would still model a number of situations that we regularly face in real life. In what follows therefore I will bracket such issues and focus on the main disagreement concerning normativity. Are conventions supported by external norms only, as Lewis claimed, or is there an intrinsic "ought" of convention? Is there a genuine distinction

[9] See e.g. Bardsley (2007) and Roth (2010) for critical analysis and overview. Notice that some 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. Tuomela's (2007) recent account in contrast strictly associates collective intentionality with normativity.

[10] Several authors after Lewis have argued that conventions can emerge from the repeated play of various kinds of games, and that coordination problems may not be special from this point of view. See e.g. Ullmann Margalit (1977), Sugden (1986) and, for an overview, Alexander (2007, Ch. 8).

between social norms and conventions – as Bicchieri (2006) suggests – or are all social institutions intrinsically normative?

Analysis and intuition

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 reveal 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. Philosophers' tales often stretch our intuitive capacities to the breaking point (as in the quoted paragraph) where clear intuitions are hard to come by. [11]

Of course this conceptual gymnastic is far from uninteresting. In telling us what a convention *really* is, a philosopher may build a complex conceptual structure that is

[11] 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 recent work in so-called "experimental philosophy" see also Knobe and Nichols (eds. 2008).

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.[12] But then agreement with linguistic practice or with intuitions in highly fictional scenarios cannot be the ultimate test of validity for philosophical reconstructions of folk concepts.

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 can we 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 analysing our folk notion of convention.

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

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 of appraisal for his theory:

> 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' general philosophy. One of Lewis' lasting contributions consists in articulating a method of philosophical analysis (the "Carnap-Ramsey-Lewis" method) that is applicable to a wide range of theories and domains – from psychology, to mathematics, colours and even holes. So readers may be misled to think that the project pursued in *Convention* is analogous to the analyses that Lewis provides elsewhere. But this is doubtful, and the best way to see it is to try 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 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.[13] This conjunction will include problematic, "Theoretical" (T) terms (mental states, for example), and unproblematic "Old" (O) terms referring to familiar objects and phenomena (facial expressions, linguistic utterances, etc.). Following Carnap, Lewis proposes that the meaning of the T-terms be defined by their function 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 "Ramseyfying" 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 theory is achieved by a tricky balancing act between general principles and intuitions about specific cases.

Frank Jackson (1998) has argued that conceptual analysis is instrumental to the goals of "serious metaphysics". The Ramseyfication of a theory, in other words, is not 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

[13] I am simplifying here (and elsewhere) for ease of presentation. See Lewis (1970, 1972) for the full account. Jackson (1998) offers a book-length exposition and defense of the analytical method that owes much to Lewis' work.

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 specific mental states.

One of Lewis' examples may help here. Imagine you are a detective in a classic Agatha Christie novel. During the investigation, you have accumulated some information (Lewis' "platitudes") that may lead to the discovery of the murderer: "the killer's accomplice has opened the door to the study room at 9.15 AM", "another accomplice has introduced the pistol between 9.30 and 10 AM", "Professor Brown (the victim) was killed around 10.10 AM", and so forth. By substituting problematic T-terms ("first accomplice", "second accomplice", "killer") with variables (X, Y, Z) we obtain a Ramsey sentence such as: "There are three individuals X, Y, Z such that X opened the door of the study room, Y introduced the pistol, and Z used it to kill Professor Brown in such and such a way". At this point we do not know yet who these individuals are, but we know the roles they have played in this murder. The substitution of variables with names (Mr White, Dr Black, and Miss Green) takes place at the fourth stage of the Carnap-Ramsey-Lewis method, corresponding to the reduction of the folk theory to the base scientific theory. The latter must be grounded on an independent empirical basis, for example on the fact that Mr White had a key to the study room, Dr Black owned a pistol, and Miss Green had an excellent motive to eliminate Professor Brown.

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 say that the principles of folk psychology are common knowledge (albeit of the tacit kind) that only requires to be made explicit for all the folk to recognize its validity. 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. All we have to do is wait for science to find entities that fit the empty boxes.

Lewis follows more or less the same strategy in his work on colours and the foundations of mathematics,[14] but the case of conventions is more complicated. Unlike folk psychology, vernacular social ontology is hardly common knowledge among the folk (cf. Lewis 1969, p. 3: "perhaps not all of us share one clear concept of convention"). In fact, if social psychologists are right we should expect the folk theory to be mistaken on a number of issues, and in a systematic way too.[15] If our intuitions are unreliable, major revisions will almost certainly be required in light of the discoveries of science.

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. 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. But if the theory is not just a conjunction of platitudes, it may well be false. Lewis Conventions may be intrinsically normative 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

[14] According to Nolan (2005).

[15] See for example Higgins and Bargh (1987), Uhlmann et al. (2008).

(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 them. But this does 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 for the reasons just stated: Lewis does not assume common knowledge of social ontology, and introduces problematic T-terms right from the start.[16] 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 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 analysis, then the issue of whether we have good intuitions about, say, normativity, would not even arise. We would not have to choose between a scientific and a folk theory, if the scientific theory was imperfect or even plainly false. This is why, according to this reading, Lewis' theory must be assessed in the laboratory, rather than in the philosopher's armchair.

[16] 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 believe that concept analysis is at best a secondary goal for Lewis (1969).

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 throws 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 about normativity? Do instrumental rationality and external norms provide an exhaustive account 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.[17] Whatever convention evolved in the early stages of the game (Red or 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 penalizes the other group members.

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

Table 2: Incentive to deviate in the tenth 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

[17] The payoffs in the tenth round are higher than in previous rounds to ensure that the decision is adequately incentivized.

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).

Notice that the payoff structure of round ten removes at least one of the extrinsic motives that support Lewis Conventions: prudential reasons (the instrumental "ought" of convention) now prescribe that one should deviate from the established regularity. But the other extrinsic motive is also under threat. Recall that according to Lewis one ought to do what answers to others' preferences, *other things being equal* – or if no countervailing reason is in place (cf. Lewis 1969, pp. 97-8). If there are prudential reasons to deviate, the ceteris paribus clause may go unfulfilled.

As a matter of fact in the laboratory less than one-third (29%) of the potential deviants decide to breach the convention. This may sound remarkably low, but it is consistent with a substantial body of experimental data.[18] So why do people decide to conform? After the tenth round the game is over and the three players will never meet again. So a purely consequentialist, forward-looking agent should not be afraid of disrupting the convention that has emerged in the previous rounds.

The simplest explanation appeals to heuristics. The resilience of conventions may be just a matter of habit, and it would disappear if the game was repeated long enough. There are however at least two reasons to discard this explanation. First, it is well known that conventions are fragile to minor changes in payoffs (Crawford et al. 2008). Subjects react to incentives, and do not mindlessly follow conventions as "fast and frugal" heuristics. Second, we know that subjects do not learn to deviate when the task is repeated. In an experiment with four "special" rounds like those of Table 2, the difference between conformity in the first and fourth special rounds was not statistically significant (see

[18] See Guala and Mittone (2010) for a more detailed discussion of the experimental results, as well as footnote 21 below.

Hodgson et al. 2012). The costs agents are willing to bear seem significantly higher than what would be justified by a bounded rationality explanation.

A *norm*-based explanation seems much more plausible. But what is a norm, and how can we detect it experimentally? When we say that "you ought to do X", we mean that we expect you to do it even if you may have some reason to the contrary. The normative element is supposed to give you an extra motivation that trumps other considerations – especially prudential reasons – that may possibly arise. (If I believe that you ought to return the money I lent you, for example, I will not take the explanation that you prefer to keep it as an adequate excuse.)

If norms trump other reasons, then conformity to a norm implies willingness to bear some costs. A player under normative pressure must be willing to give up something to conform to the established rule. In our experimental setting, 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.[19] Our tenth round can then be used as an "acid test" to detect the influence of normative forces.

The acid test must be handled with care however. In principle a potential deviant may feel obliged to conform to a convention even though she believes that the other players only have plain (non-normative) expectations regarding her future choices. The deviant might feel guilty because she realizes that her choice will affect the payoffs of the other players. And of course many of us *do* care about others' payoffs. These other-regarding

[19] This use of monetary incentives is very common in experimental psychology and economics. In experiments with Prisoner's Dilemma or Ultimatum games, for example, monetary incentives are used to detect factors that prompt individuals to deviate from narrow selfish behaviour (maximization of one's material gain). By observing deviations from narrow selfishness we can try to reconstruct agents' 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 Binmore (1994), Bicchieri (2006), Guala (2006), and Woodward (2009).

inclinations are governed by independent norms that ought to be carefully distinguished from the intrinsic normativity of convention that our experiment intends to capture. [20]

A norm of *altruism* ("you ought to help or at least not cause harm to the members of your group") would 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 her personal gain. Similarly, norms of *fairness or equality* may prescribe to conform to the behaviour of other group members because this is the way to achieve an equal distribution of the resources.

How can we separate these "external" norms of altruism, fairness or equality from the intrinsic normative force of convention? We conjectured earlier that the decision to conform in the tenth round may be influenced by the history of play that has developed in one's own group. Repeated team play may generate normative expectations of conformity, independently from the considerations of fairness discussed above.

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 should subtract the intrinsic force of convention, in other words, and leave only the effect of external norms. This is what happens in the one-shot game represented in Figure 2.

[20] I am speaking of norms for simplicity here. The same behaviour can be explained by theories of other-regarding preferences with deeper (perhaps biological, innate) roots. 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).

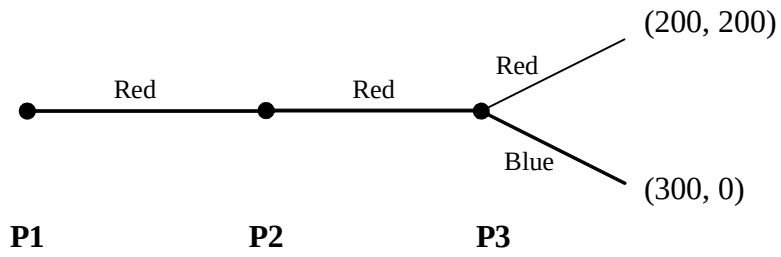

Figure 2: The one-shot game.

We analyse the game from the viewpoint of the potential deviant ("Player 3"). Suppose the first two players have played Red. The potential deviant can observe their moves, and then decide whether to choose the same colour or not. Notice that at this point she is facing exactly the same decision situation as in 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 behaviour of the potential deviant, 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, however, 68% of the experimental subjects decide to deviate, compared to 29% in the repeated game.[21] The mere fact of playing together for nine rounds is sufficient to enhance conformity.

[21] This replicates the results of other 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 once we compare these results with those from another condition where 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 34% of the subjects towards the equitable outcome.

Conventions are not only sustained by external norms of cooperation, but also by an intrinsic normative pressure to conform to an established regularity.

The normativity of 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 to pursue her own self-interest, and to avoid damaging other players without cause. But another reason motivates real players facing simple choices in laboratory settings. When players build a history of joint action, they unintendedly create an additional pressure towards conformity that goes beyond the "ought" of individual rationality and other-regarding external norms. Whether these additional normative expectations are to be explained by a joint commitment (Gilbert 1989) or some other mechanism (Bicchieri 2006) 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 family members, friends, and colleagues.

The intrinsic normativity of Lewis Conventions has been noted before by philosophers interested in the analysis of folk concepts. As I have argued in this paper, however, there are good reasons to believe that Lewis was not primarily analysing a folk theory. If this is true, then his theory should not be appraised using criteria that are appropriate for the analysis of folk concepts. The relevant criteria are *scientific*, and Lewis' theory should be

assessed in the light of these only. Intuitions do play a role in the development of scientific theories, but they are not the evidence against which such theories are tested. They rather work as heuristic devices, suggesting mechanisms and hypotheses which must then be investigated empirically.

In this paper I have reported the results of experiments that attempt to do that. The data suggest that Lewis Conventions tend to acquire intrinsic normative force through repeated play, and any future model will have to account for these results. Their implications are non-trivial, both in the theoretical and practical realm. Here I will just mention one: the intrinsic normativity of conventions implies that habits and customs may be more difficult to disrupt than a pure rational choice analysis suggests. While individuals do react to incentives, they also display a remarkable reluctance to abandon traditional equilibria once they are in place. Experimental evidence has important implications for social policy, social ontology, and political philosophy alike.

References

Alexander, J.M. (2007). *The structural evolution of morality*. Cambridge: Cambridge University Press.

Armel, K.C., Beaumel, A. & Rangel, A. (2008). Biasing simple choices by manipulating relative visual attention. *Judgment and Decision Making*, 3, 396–403.

Bacharach, M. (2006). *Beyond individual choice: Teams and frames in game theory*. Princeton: Princeton University Press.

Bardsley, N. (2007). On collective intentions: Collective action in economics and philosophy. *Synthese*, 157, 141-159.

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

Binmore, K. (1994). *Game theory and the social contract, vol. 1: Playing fair*. Cambridge Mass.: MIT Press.

Bratman, M. (1993). Shared intention. *Ethics*, 104, 97-113.

Carrara, M. & Varzi, A.C. (2001). Ontological commitment and reconstructivism. *Erkenntnis*, 55, 33-50.

Charness, G. & Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117, 817-69.

Crawford, V.P., Gneezy, U., & Rottenstreich, Y. (2008). The power of focal points is limited: even minute payoff asymmetry may yield large coordination failures. *American Economic Review*, 98, 1443-1458.

Gilbert, M. (1989). *On social facts*. London: Routledge.

Gilbert, M. (2008). Social convention revisited. *Topoi*, 27, 5-16.

Guala, F. (2006). Has game theory been refuted? *Journal of Philosophy*, 103, 239-63.

Guala, F. & Mittone, L. (2008). An experimental study of conventions and norms. CEEL Working Paper 8-10, University of Trento.

Guala, F. & Mittone, L. (2010). How history and convention create norms: an experimental study. *Journal of Economic Psychology*, 31, 749-756.

Higgins, E.T. & Bargh, J.A. (1987). Social cognition and social perception. *Annual Review of Psychology*, 38, 369-425.

Hodgson, T., Guala, F., Miller, T. & Summers, I. (2012). Limbic and prefrontal activity during conformity and violation of norms in a coordination game. *Journal of Neuroscience, Psychology, and Economics*, 5, 1-17.

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. & Nichols, S. (eds. 2008). *Experimental philosophy*. New York: Oxford University Press.

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.

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

Roth, A.S. (2010). Shared agency. In Zalta, E.N. (ed.) *Stanford encyclopedia of philosophy*. http://plato.stanford.edu/entries/shared-agency/

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. & Pollack, M.E. (eds.) *Intentions in communication*. Cambridge, Mass.: MIT Press.

Skyrms, B. (1996). *Evolution of the social contract*. New York: Cambridge University 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. (2002). *The philosophy of social practices*. Cambridge: Cambridge University Press.

Tuomela, R. (2007). *The philosophy of sociality: The shared point of view*. Oxford: Oxford University Press.

Uhlmann, E.L., Pizarro, D.A.& Bloom, P. (2008). Varieties of social cognition. *Journal for the Theory of Social Behaviour*, 38, 293-322.

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

Woodward, J. (2009). Experimental investigations of social preferences. In Kincaid, H. and Ross, D. (eds.) *The Oxford handbook of the philosophy of economics*. New York: Oxford University Press.

Appendix: experimental procedures

The experiments described in this paper were run at the universities of Exeter and Trento, using the typical procedures of experimental economics. Subjects were recruited using email lists from the population of graduate and undergraduate students. Volunteers registered for one of the sessions and, as they arrived at the lab, were seated randomly at 18 computer terminals separated by partitions. After signing a consent form, they were asked to read the experimental instructions illustrating the main features of the task. Each subject participated in one condition only, and all comparisons took place across subjects. In all conditions subjects played in groups of three players, with random selection of group membership, anonymity, and without communication. They received a show-up fee of three Euro, and on top of that received whatever they earned in the experimental task. Individual payoffs were calculated in terms of "experimental tokens" which were converted into real money at the exchange rate of three cents per token (or one Euro = 33 tokens).

Subjects were told in the instructions that the game would last for ten rounds only. They were also told that the payoffs could change during the course of the game, but that all players would be informed in advance if this happened. In the instructions no specific details were provided regarding the payoff structure of these "special rounds". As a matter of fact, in the tenth and last round all groups faced a "temptation" game with the payoff structure represented in Table 2.

One of the key experimental issues was how to control for the effect of trust and reciprocity in the repeated and in the one-shot conditions. To this purpose we compared a condition where the potential deviant observed the moves of the other players, with a condition where the moves of players 1 and 2 were controlled by a computer that enforced the convention that had emerged in earlier rounds (because their move was unintentional, player 3 could not reciprocate). Since there is no difference between these two versions of the experiment, we conclude that trust and reciprocity do not play a major role in the resilience of social conventions. More details about the experimental procedures and statistical data-analysis can be found in Guala and Mittone (2008, 2010).