Topoi (2008) 27:73–86
DOI 10.1007/s11245-008-9029-0

Conventions and Moral Norms: The Legacy of Lewis

Bruno Verbeek

Published online: 1 July 2008

Abstract David Lewis' *Convention* has been a major source of inspiration for philosophers and social scientists alike for the analysis of norms. In this essay, I demonstrate its usefulness for the analysis of some moral norms. At the same time, conventionalism with regards to moral norms has attracted sustained criticism. I discuss three major strands of criticism and propose how these can be met. First, I discuss the criticism that Lewis conventions analyze norms in situations with no conflict of interest, whereas most, if not all, moral norms deal with situations with conflicting interests. This criticism can be answered by showing that conventions can emerge in those contexts as well. Secondly, I discuss the objection that this type of conventionalism, inspired by Lewis, presents moral norms as fundamentally contingent, whereas most, if not all, moral norms are not. However, such critics fail to appreciate that conventions are not radically contingent. Moreover, if one distinguishes the question as to why an individual should comply with a norm from the question whether the norm in question itself can be justified, a core element of the complaint of contingency disappears. The third objection to conventionalism concerns the way in which conventionalists justify norms. I argue that reflection upon the way in which according to Lewis norms are justified reveals a fundamental tension in his theory. Possible solutions to this tension all have in common that the complaint of contingency returns in some form. Therefore, this third complaint cannot be avoided altogether.

B. Verbeek (✉)
Department of Philosophy, University of Leiden, Postbus 9515, 2300 Leiden, RA, The Netherlands
e-mail: B.Verbeek@phil.leidenuniv.nl

Keywords Conventions · Norms · David Lewis · Conventionalism

1 Introduction

In 1969, David Lewis published his seminal work *Convention*. It has been an influential work in many areas of philosophy, and it has had a major impact on, at least some, moral philosophers.[1] The reason for this influence is easy to understand. *Convention* strongly suggests a model for a naturalistic account of social norms in the tradition of Hume.[2] It showed how norms can emerge and how they can have genuine normativity for those participating in them, without the necessity of invoking normative facts or other metaphysically suspicious entities.

Conventionalism, as I will call the body of thought inspired by Lewis, has tried to apply Lewis' analysis of linguistic conventions to social norms, most notably, moral norms. The task for conventionalists is to investigate which features of moral norms can be explained with the analysis. Conventionalists vary in how strong they take the analysis to be. There are those who insist that almost all aspects of all moral norms can be treated as complex social conventions.[3] Others take a more moderate view and argue that Lewis' analysis is helpful in explaining some features of some norms, but not all moral norms.[4] Yet others have

[1] Nevertheless, it took quite a bit of time before the relevance of *Convention* for moral philosophy was fully recognized. See also Verbeek (2002a).

[2] Hume (1998, 2001).

[3] E.g., Skyrms (1996, postscript) comes close to defending this position.

[4] E.g., Sugden (1986); Den Hartogh (2002); Bicchieri (2006); Verbeek (2007).

used the analysis to make finer grained distinctions between kinds of social and moral norms.[5] In so far as moral norms are concerned, most conventionalists agree that the rules of what Hume calls 'justice' in the *Treatise* are particularly suitable candidates for a conventionalist analysis, whereas questions of 'broad morality' typically fall out of the scope of the generalized Lewisian project.[6]

At the same time, conventionalism has had many detractors. Moral norms, so these critics argued, cannot be analyzed analogous to the sort of linguistic conventions that Lewis described. Three differences seem particularly significant. First, whereas linguistic conventions are supposed to cover situations where there is no conflict of interest, moral norms guide us especially in situations where there is considerable conflict. Secondly, linguistic conventions are just that: conventions. It is just a matter of convention whether to refer to members of the species of the genus *Equus* as 'horses' rather than 'chevaux'. Linguistic relativism—the notion that linguistic rules vary according to place, time and history—therefore, is to be expected (and indeed the case). However, moral norms, like those requiring people to refrain from murder, for example, are anything but conventional. Such norms, many argue, are universal and categorical, rather than peculiar to our society and relative to our practices. Finally, there is the objection that conventions cannot be justified in the appropriate manner. At best one can tell a causal story how a particular convention came into existence, but one cannot justify that this is the right convention within Lewis' theory. However, moral norms are proper objects of justification and moral debate. Since moral norms can be justified, it is argued, moral norms cannot be conventional rules.

In what follows, I will explain how one can answer these three objections in a Lewisian spirit. In doing so both the merits as well as the limits of this way of thinking of morality will become clearer. First, I will rehearse some of the main steps in Lewis' analysis of conventions. Next, applications of these steps to a situation with more conflict of interest will be introduced. This suggests a principled reply to the first type of objections. I then turn to the second criticism that moral norms, unlike conventions, are not contingent. I will argue that there is a principled response available for the defender of Lewis. However, this response turns out to be something of a poison pill, since it points to an inherent dilemma for Lewis' analysis. I conclude with some remarks about the merits of Lewis' project for moral philosophy in spite of this dilemma.

[5] E.g., Elster (2007, pp. 353–371).

[6] Hume (2001, III, 2, sections ii–x). The distinction between 'broad morality', i.e., those rules of conduct that guide one's entire life, and 'narrow morality', i.e., the more limited rules of conduct that enable people to coordinate and cooperate in society comes from Mackie (1977, ch. 5).

2 Lewis conventions

Suppose you and I are to meet each other. For some reason we did not settle upon a place where we would meet. It is completely indifferent where we meet, as long as we meet each other. Where would you go? Since our only concern is to meet, we are engaged in a pure coordination problem (see Table 1).[7]

Table 1 A pure coordination problem

	Location 1	Location 2	Location 3	Etc.
Location 1	1, 1	0, 0	0, 0	0, 0
Location 2	0, 0	1, 1	0, 0	0, 0
Location 3	0, 0	0, 0	1, 1	0, 0
Etc.	0, 0	0, 0	0, 0	1, 1

Standard game theory, at least the sort that was around when Lewis wrote his dissertation, is unhelpful here. First, notice that neither you nor I have any independent reason to go to any of these locations. For neither of us, it is the case that we have reason to go to any of the locations regardless of what the other does. In the parlance of game theory, there is no dominant strategy for either one of us, nor is there any dominated strategy. In determining where to go, each of us needs to anticipate the choices of the other. Secondly, our reasons for going to any of these locations refer to each other. What reasons are there for my going to location 1? I have such a reason only if I believe that you will go to location 1. Why would I believe that? Well only if I believe that you believe that I will go to location 1. That is, only if I believe that you have a reason to go to location 1. What reason do I have for that belief? I have a reason for this belief, if I believe that you believe that I believe that you believe that I will go to location 1. In other words, I have such a reason if I believe that you have a reason to believe that I have a reason to go to location 1. My reasons for going to location 1 depend on your reasons for going to location 1 and *vice versa*. Our reasons are interdependent. Consequently, any choice of location looks arbitrary from the individual's perspective.[8]

[7] A pure coordination problem is a situation of interdependent decision by two or more agents in which there are multiple proper equilibria consisting of corresponding strategies, where none of the agents has a preference for any of these equilibria. And an equilibrium is 'proper' if, for all agents, the equilibrium outcome is preferable over all other feasible outcomes given the choices of the other agents. See also Lewis (1969, pp. 14–32).

[8] Notice that this interdependency is brought about by the existence of multiple Nash equilibria and the absence of any dominant strategy. An outcome is a Nash equilibrium if for each individual it is the case that they could not have improved their outcomes given the choice of all other individuals. A strategy is dominant if and only if it is always gives at least as good an outcome as any other strategy. For precise definitions, see Fudenberg and Tirole (1992, sections 1.1–1.2).

However, in real life, people often have no problem to coordinate in such situations. In the late 1950s, Thomas Schelling presented subjects with the problem of meeting in New York and a surprisingly large number responded that they would go to Grand Central Station in such a case.[9] As is well known, Schelling's explanation for this observation was that real people—as opposed to the idealized agents of game theory—somehow are sensitive to information external to the formal description of the game.[10] They regard one of the possible equilibria as *salient*.[11] What exactly salience is in this context was relatively broad. Schelling suggested that it could be some form of psychological prominence. For example, the human eye is most sensitive to the color red. Therefore, when asked to coordinate on a color ("You will each get a reward if you pick the same color as the other"), the color red is prominent. However, there are other forms of salience as well which cannot be explained so easily by reference to psychological prominence. Lewis points out that explicit agreement and—what is most germane in our context—precedent are forms of salience as well.[12]

Be that as it may, what is noteworthy about Schelling's explanation is that on his view successful coordination is a-rational. Standard game-theoretic reasoning does not recommend any pure strategy over another on the traditional picture even though coordination is preferable over the failure of coordination.[13]

Lewis' theory of conventions provides a way to allow such external information to play a role in reasoning about one's choice of strategy. The basic idea is presented in the first two chapters of *Convention*. Suppose that I believe, with some sufficient degree of certainty, that you have a tendency to go to location 1, e.g., because you find it salient (e.g., you have gone there several times in the past in situations just like these). Subsequently, I can form the belief that you believe that I believe that you will go to location 1. That is, I now have arrived at an additional reason to believe that you will go to location 1. If that is not enough, if the weight of these reasons is not sufficient for me to go to location 1, I can form a belief of yet a higher order and add to the balance of reasons. At some point, my first order belief about your tendency to go to location 1 in combination with these higher-order beliefs will be sufficient reason for me to go to location 1.[14] In other words, what Lewis argued is that higher-order beliefs can add to the weight of reasons for a choice of strategy. The formation of these higher-order beliefs is justified according to Lewis if the salience of location 1, as well as the structure of the game and that we are rational, is common knowledge between us.[15] That is, because I know that you know that I know … that location 1 is salient, I can infer that you believe that I believe that you will go to location 1—and similarly for all higher-order beliefs.

These two ideas, the notion that higher-order reasons can add to the weight of first order reasons and the claim that salience can be a first-order reason when it is object of common knowledge are the fundamental improvements upon Schelling's explanation. With these additions, it seems that the individuals in Schelling's experiments were far from a-rational. Instead, they used commonly known clues about each other's tendencies to reason correctly to a choice of strategy.

Suppose the members of a group manage to coordinate successfully on an occasion. There now is a precedent. Suppose that members of the same group encounter a similar coordination problem. Lewis suggests that the precedent will make one of the coordination equilibria salient, thus reinforcing the tendency of agents to do their

[9] Schelling (1960, p.58). His respondents were all students at Yale in the late 1950's. Grand Central Station was the place where most of them would arrive in New York.

[10] The formal description of a game consists of three elements. The (finite) set of individual players $i \in I$; a pure strategy space for each individual S_i and payoff functions u_i that assigns an individual i a utility for each strategy profile $s = (s1, s2, s3, \ldots)$. All information about the strategy labels the individuals use (e.g., 'location 1' or 'Grand Central Station'), or further characterizations of the strategy profiles (e.g., 'we all pick red', or 'we all go to Grand Central Station') beyond these three elements is excluded from the formal description of a game. Schelling's point is that agents use such excluded information to coordinate their actions.

[11] Strictly speaking, only combinations of strategies are equilibria. Outcomes are not equilibria, though they can be the result of equilibrium strategies. In this essay, I will be less strict. Sometimes I will refer to strategies as an equilibrium and sometimes I will refer to an outcome as an equilibrium. It will be clear in the context which of these two usages are meant. Nothing important depends on this—admittedly sloppy—use of terminology. Thanks to an anonymous referee for pointing this out to me.

[12] Lewis (1969, pp. 33–41); Mehta et al. (1994a; b) demonstrate that there are forms of salience that are intrinsically conventional. See also Postema's contribution to this volume.

[13] One referee reminded me that standard game theory in this case recommends following a mixed strategy: agents are recommended to randomize over the available strategies with equal probabilities. Randomization does not ensure that coordination will result. What is more, mixed strategies are suspicious as rational recommendation in coordination games in any case since if others follow the equilibrium mixed strategy I can expect the same pay-off regardless which strategy I follow.

[14] A first-order belief is a normal belief about some state of affairs; a second-order belief is a belief about a first-order belief; ...; an n-order belief is a belief about a belief of the order n-1. Lewis (1969, pp. 28–32).

[15] To be precise, in addition to common knowledge of the salience of location 1, we need to have common knowledge of the nature of our predicament (that we want to coordinate) and of our rationality (to allow for robust predictions of each other's conclusions) as well as our inductive standards and background information (Lewis 1969, p. 52–56).

part to realize this outcome.[16] If members of a group regularly coordinate successfully, they will start to notice this regularity. Consequently, they will develop a conditional preference for conforming to this regularity. Members typically prefer to conform to the regularity if others do so as well. Since there is this regularity, a general expectation that people will conform is formed. Consequently, a stable pattern of behavior emerges that is based on the general expectation of each group member about the typical way that others will behave. At that point a convention has emerged. If people like you and me often miss each other in New York City, we will go to Grand Central Station. It will be the convention to go there in such cases. Lewis' famous definition of a convention is the following:

> A regularity R in the behavior of members of a population P when they are agents in a recurrent situation S is a convention if and only if it is true that, and it is common knowledge in P that, in almost any instance of S among members of P,
>
> (1) almost everyone conforms to R;
> (2) almost everyone expects almost everyone else to conform to R;
> (3) almost everyone has approximately the same preferences regarding all possible combinations of actions;
> (4) almost everyone prefers that any one more conform to R, on condition that almost everyone conform to R;
> (5) almost everyone would prefer that anyone more conform to R' on condition that almost everyone conform to R';[17]
>
> where R' is some possible regularity in the behavior of members of P in S, such that almost no one in almost any instance of S among members of P could conform both to R' and to R (Lewis 1969, p. 78).

Thus, there is a convention to meet each other in Grand Central Station if (1) we, and others like us, meet each other in the situation of Table 1 in Grand Central Station and we do so regularly, because (2) we expect each other to do so (for example, because we have met here the last couple of times), and (3) we all have an interest in meeting each other, and, (4) even though we are indifferent where we will meet, we would prefer that we meet in Grand Central Station, on the condition that almost everyone would go to Grand Central Station in such cases and (5) it is the case that we, and all others like us, would prefer to meet on, say, Times Square, if that is where people would go when they loose each other in New York City.

Lewis' analysis of conventions then achieves something remarkable. It shows, first, that conventions help to solve coordination problems. Secondly, it shows that conventions emerge in recurring coordination problems. Thirdly, it avoids any dubious functionalism. Conventions do not emerge because they fulfill a beneficial function for a population. Instead, Lewis provides an explanation which firmly bases both the answer as to why conventions emerge as well as why people comply with these conventions in a theory that does justice to individual intentionality and rationality.

These features explain the attractiveness of the suggestion that moral norms could be analyzed as conventions. A moral agent is a person who acts as morality dictates him. He does not merely conform to the requirements of morality. An agent could conform for reasons that have nothing to do with the fact that an action is required. In contrast, the reasons of a moral agent for acting morally are precisely that his actions are required by morality. That is, a moral agent complies with morality. For example, a moral agent does not refrain from cheating others simply because it is inconvenient or because it would ruin his reputation. A moral agent does not cheat others because he should not. This is a central feature of moral norms and it is the task of ethical theory to explain this feature. This is a difficult task because compliance seems problematic from the point of view of rationality. Suppose that a moral norm requires the agent to φ. Then, either there are independent reasons to φ, in which case the fact that a norm requires the agent to φ is irrelevant.[18] Or, there are no independent reasons to φ, in which case it is hard to see how the mere fact that a norm requires it makes φ-ing rational. In other words, a moral norm is irrelevant for a rational agent or it is irrational.[19]

This dilemma could be avoided if one tinkered with the notion of rationality, so as to render compliance rational and vindicate the rationality of moral norms. However, for those who are reluctant to reject the instrumentalist notion of rationality as it is expressed in standard rational choice theory, conventionalism suggests another way of avoiding the dilemma. Consider the reasons of an agent to go to location 1. There are no independent reasons that tell him to go to any of these locations in particular. The only independent reason there is to meet with the other agent. However, the existence of a convention gives the agent reasons to go to location 1. The reason of our agent to go to location 1 is that in those situations there is a convention to go to location 1. The agent, therefore, complies with the convention in going to location 1. Such compliance is not

[16] Lewis (1969, pp. 36–42).

[17] Note that (5) requires that there be multiple equilibria.

[18] Independent, that is, from the fact that the norm requires the agent to φ.

[19] McClennen and Shapiro (1998).

irrational, nor is the convention irrelevant for the rationality of going to location 1. So, if a similar story could be constructed for other norms one could vindicate the rationality of compliance with these norms.

3 A Refinement of Lewis' Definition

Before considering whether Lewis' analysis of conventions is applicable to other, especially moral, contexts, we need to make two modifications on Lewis' definition. A convention, on Lewis definition, is a species of convergence in behavior. However, such a convergence in behavior often is a sign that agents are following a norm, but it is not itself a norm. A (temporary) disruption of the norm is not sufficient for the non-existence of a norm. Norms can be violated after all—even on a relatively large scale. Furthermore, there can be all kinds of reasons why the behavior of agents converges. Some subjects in Schelling's experiment may have preferred to go to Grand Central Station because they admired the architecture. Others may have opted for Grand Central Station without any thought; sheer habit may have led them to go there. For these reasons it is probably better to think of conventions as (part of) the (justifying) reason for agents to display such converging behavior.[20] Consequently, we will have to formulate Lewis' definition of a convention as expectations about behavior rather than actual behavior.

This modification is in the spirit of Lewis' analysis as he calls attention to the mental states of the agents conforming to a convention and makes these part of his definition. The emphasis on the agents' preferences and their beliefs, especially, their common knowledge, simply is an emphasis on the reasons of the agents for conforming to the convention.

The second point we need to make is about condition (3) which requires that all agents have similar preferences for all combinations of outcomes. This seems overly restrictive. Lewis' focus was on linguistic rules where the underlying interests of all concerned resemble (pure) coordination problems. However, if one wishes to extend the analysis to other types of norms, including moral ones, and other types of interaction problems, we need to drop this condition. So from here on, I will assume that this is not a necessary condition for the existence of a convention.

4 Conventions in Situations of Conflicting Interest

It might be argued that regularities in behavior modeled after Lewis conventions are not relevant for ethics. One reason for such sentiments might be the following. Moral norms have a 'point'. They are invoked of situations of potential conflict of interest. They help avoiding threatening sub-optimality. Edna Ullmann-Margalit has argued that this is the *raison d'être* of moral norms.[21] The most notorious of such situations is, without doubt, the prisoner's dilemma. Ullmann-Margalit argued that moral norms are commitment devices that prevent agents from 'cheating' or 'free riding' in prisoner's dilemmas. The prisoner's dilemma falls outside of the scope of Lewis' definition of a convention, because the single equilibrium in this situation (therefore (5) is not met) is the result of the dominant strategies of the agents. That is to say, the equilibrium is such that agents have reason to choose their equilibrium strategy regardless whether others do so as well. Therefore, they need not expect others to conform as well (as required by (2)), nor is it the case that the agents prefer anyone to conform to their equilibrium strategies, provided others do so as well (as required by (4)). If anything, agents would prefer others not to choose the equilibrium strategy as this opens up the possibility of 'cheating'.

However, Lewis' analysis can be used to demonstrate that conventions in prisoner's dilemmas are solutions to a coordination problem as well. Furthermore, with a few adaptations, Lewis' analysis can be used to demonstrate that in a prisoner's dilemma conventions focusing on the Pareto-optimal outcome can emerge. Finally, I will suggest that some of these conventions are moral norms, thus vindicating my suggestion in the introduction that at least some moral norms are conventional.

Consider Hume's famous example of the two farmers who are considering whether to help each other reap their harvests:

> Your corn is ripe to-day; mine will be so tomorrow. 'Tis profitable for us both, that I shou'd labour with you to-day, and that you shou'd aid me to-morrow. I have no kindness for you, and know you have as little for me. I will not, therefore, take any pains upon your account; and shou'd I labour with you upon my own account, in expectation of a return, I know I shou'd be disappointed, and that I shou'd in vain depend upon your gratitude. Here then I leave you to labour alone: You treat me in the same manner. The seasons change; and both of us lose our harvests for want of mutual confidence and security. (Hume 2001, III, 2, v)

If each attempts to harvest their corn without help, they will not be able to bring in all the corn in time before the storm. If they help each other, they can work more

[20] I argue for this in more detail in Verbeek (2007). See also Den Hartogh (2002).

[21] Ullmann-Margalit (1977). For similar views, see Mackie (1977).

efficiently and both their harvests are saved. Toiling the land, however, is hard work and both would prefer help with their own harvest but not to come to the aid of the other.

I will ignore the aspect of timing that is clearly present in Hume's example. I pretend that the farmers somehow help each other simultaneously.[22] Then we have a standard two-person prisoner's dilemma:

Table 2 Prisoner's dilemma

		B	
		Mutual aid	No aid
A	Mutual aid	(2, 2)	(0, 3)
	No aid	(3, 0)	(1, 1)

The farmers could promise to help each other, but why would they honor such a promise in the absence of external enforcement? For with a promise it still is the case that the dominant strategy for A and B is not to provide aid to the other. Perhaps it could be argued that it is morally required that they both assist each other. Again, why would A and B comply with such a norm? However, if we can show that in such situations there could be a convention to assist each other, the idea that A and B should comply with the conventional norm to assist each other is not so strange. It would lend credence to conventionalism.

If this is a situation that occurs only once, a convention cannot emerge. Both A and B will not help each other, thus realizing a sub-optimal outcome. This is their dominant choice. In this case there is no convention possible as we saw above. So it seems as if Lewis conventions can emerge only in contexts of (pure) coordination problems. However, this is too fast. In order to see that even in prisoner's dilemmas conventions of mutual aid can emerge, let us introduce some plausible assumptions about the context of the problem for the farmers.

First, I will assume that this situation arises more often than once. There are many situations like this occurring within this community. That is, this is a repeated prisoner's dilemma.[23] It is unrealistic to assume that agents in such a population will meet each other in an infinite number of occasions—after all, as the song goes, 'we all have to go sometime'. However, few of us, if any, know when. So, let us assume that when two agents from this community are engaged in a prisoner's dilemma, there is a probability p ($0 \leq p < 1$) that they will meet each other again in a similar situation. For example, in a community of farmers in which A and B have the option for mutual aid, there will be a new season with new harvests and new opportunities for mutual aid. However, there is no guarantee that A and B will be in this situation again. For example, A or B (or, indeed, both) may give up farming and leave the community.[24] In such a repeated game, there is a plethora of possible strategies that could be followed by each of the agents. In order to facilitate the analysis I will restrict the class of repeated prisoner's dilemmas to that class where strategies other than 'never assist' could emerge as well. Therefore, the third assumption that I will make is that each member of the community stands to gain something by cooperating and giving mutual aid. This is only the case if the average number of rounds is large enough. To this end, I will assume that $p > {}^1/_3$. This means that the average number of times n that A and B will be able to help each other with the harvest in this manner is 2 or more.[25] If $p \leq {}^1/_3$, the average number of rounds equals 1 and the agents are in fact playing one-shot prisoner's dilemmas with each other and no convention could emerge.

Finally, I have to introduce an assumption that Lewis does not make, but it seems relatively harmless in the present context. Individuals tend to remember the actions of others with whom they interacted. In other words, people can form a reputation in this context.[26]

If these four assumptions are made, it looks like all kinds of cooperative strategies could emerge. Perhaps the most famous among them is *tit-for-tat* (TFT).[27] TFT starts out with cooperation and repeats the move of the other player in each subsequent round. Let n be the number of the round, then TFT plays cooperate in $n = 1$ and in each subsequent round $n > 1$, TFT repeats the move of the opponent in n-1. TFT is an equilibrium in the repeated finite prisoner's dilemma of which the repeated harvest conundrum is an example. That is, TFT is a strategy such that if it is followed in this population, neither A nor B

[22] Gauthier (1986, ch. 6) and McClennen (1990) have argued, on independent grounds, that the removal of this sequential aspect really matters. They offer arguments as to why the farmers should cooperate in a single sequential prisoner's dilemma. I ignore this complication in this paper.

[23] This assumption is in line with Lewis' own thinking as he claims that precedence is one of the most common forms of salience.

[24] Let n be the number of the rounds A and B interact. Then, there is a chance p^{n-1} that A and B will face each other again in such a situation. Since $0 < p < 1$, $p^{n-1} \to 0$ for large n.

[25] Suppose that $p = 1/3$. If A and B meet each other, the number of rounds that they will meet each other in total will be $n = 1 + 1/3 + (1/3)^2 + (1/3)^3 + \ldots + (1/3)^n \to 1.5$ Since n is an integer, it will equal 2.

[26] This is misleading, for a reputation is not the same thing as a register of past actions. See Morris (1999) on this. I ignore this issue here.

[27] See Axelrod (1981); Axelrod and Hamilton (1981); Axelrod (1984); Axelrod (1986).

could improve given the choice of the other for TFT. The proof is simple and well-known.[28] Suppose B plays TFT and A knows this. Then in round i, independent of the value of i, these two questions must have a determinate answer: (1) if B will cooperate in i, can it be part of a best reply for A to cooperate as well? (2) If B will defect in i, can it be part of a best reply for A to cooperate in that round? Suppose the answer to (1) is "Yes", then let $i = 1$. But if A cooperates in $i = 1$ then B will cooperate in $i = 2$ as well. So in $i = 2$ the situation is identical to $i = 1$. Once again it will be your best reply to cooperate in $i = 2$. Thus if the answer to (1) is "Yes" cooperating in every round is your best reply against TFT.

Suppose however the answer is "No", then any best reply to TFT must defect in $i = 1$, this ensures that B will defect in $i = 2$. Now what to do? Depending on the answer to question (2) A must defect if it is "No" and cooperate if it is "Yes". If it is "No" A's best reply to TFT is always defect. This is strategy D. If it is "Yes" B will cooperate in $i = 3$ which brings A again in the same situation as in (1) where $i = 1$, so A must defect again. A's best reply then would be to alternate between defect and cooperate starting with defection in $i = 1$. Let us call this strategy A.

We can calculate the expected utility of each of these strategies playing against TFT using the values of Table 2:

$$\text{E(TFT, TFT)} = 2(1 + p + p^2 + p^3 + ...) = 2/(1 - p)$$

$$\text{E(D, TFT)} = 3 + p + p^2 + p^3 + ... = 2 + (1/(1 - p))$$

$$\text{E(A, TFT)} = 3 + 3p^2 + 3p^4 ... = 3/(1 - p^2)$$

Since $p > {}^1/_3$, E(TFT,TFT) > E(D,TFT) and E(TFT,TFT) > E(A,TFT). This means that TFT is better than D or A against itself. But, given our answers to (1) and (2), one of these three strategies must be a best reply against TFT. Therefore, TFT is a best reply against itself. From this it follows that TFT is an equilibrium. Given that others are following TFT, there is no reason to prefer to pursue another strategy than TFT, since TFT does at least as well as any other strategy in such a population.

However, this is not enough to show that TFT is a possible Lewis convention in this population. TFT could not satisfy condition (4) of Lewis' definition. Given the assumptions it cannot be the case that each prefers that any one more conforms to TFT when almost everyone conforms to TFT. In a population of TFT players, a strategy of unconditional cooperation C ('always cooperate, no matter what the other has done in the previous round') does as well as TFT and there is no reason for TFT players to prefer others to follow TFT rather than C.[29] In game theoretic terms, TFT is not *stable*.[30]

However, suppose that the agents in this population are not perfect in the execution of their strategies.[31] They sometimes make mistakes. Now TFT is no longer an equilibrium because it is no longer a best reply to itself when a mistake has been made. Suppose two TFT players play against each other and one of them makes a mistake; he defects. In the next round his opponent will react by defecting while he himself will cooperate since his opponent cooperated in the previous round. In the following round this will be reversed, and so on. The two contestants will be locked in a cooperate-defect sequence (in other words, they play strategy A) and could end up in an "always defect" sequence if another mistake is made.[32]

Consider then the class of strategies T. A member of this class is T_1. This strategy copes with the possibility of mistakes. It implies the concept of *good standing*. A player who is in good standing is entitled to expect cooperation from his opponent. At the beginning of a sequence each player is in good standing and remains so, provided each player always cooperates when T_1 prescribes this. If any player defects though T_1 tells him to cooperate, he looses his good standing. He regains it if he cooperates unconditionally in one subsequent round: hence the name T_1. There are of course other strategies using good standing, T_2, T_3, T_4, etc., which make up the class of strategies T. These strategies demand two, three, four, or more rounds of cooperation before the other player regains good standing.

T_1 can be described as "Cooperate if your opponent is in good standing, or if you are not; otherwise defect". The only difference between TFT and T_1 is in the moves after a player has made a mistake and defected. Unlike "normal" TFT, T_1 is stable. The proof is similar to the argument used to show that TFT is an equilibrium. When you go into round i, three situations are possible: (1) both you and your

[28] See Axelrod (1981); Axelrod and Hamilton (1981).

[29] In other words, condition (4) of Lewis' definition requires that a convention is a uniquely optimal reply against itself. Let I be a strategy and E(I, J) denote the expected utility of playing I in a population that follows J. On Lewis' definition I is a convention only if E(I,I) > E(J,I) where I ≠ J. This is stronger than simply requiring that a convention be an equilibrium in a game, since I is an equilibrium only if E(I,I) ≥ E(J,I). Also, it is stronger than requiring that a convention is an evolutionary stable strategy (Smith 1982).

[30] A strategy is *stable* if and only if it either is the only best reply against itself, or if it is a best reply against itself, but a better reply against those other strategies. More precise, a strategy is stable if and only if: (1) E(I,I) ≥ E(J,J), *and* (2) either E(I,I) > E(I,J) or E(I,J) > E(J,J) (Smith 1982, p. 10; Sugden 1986, pp. 28–29).

[31] In what follows, I am taking up a suggestion of Sugden (1986).

[32] In Axelrod's computer tournament exactly such a deadlock ending in continuous mutual defection happened between TFT and the strategy called JOSS. JOSS played tit-for-tat but defected occasionally (10% of the time). One could say, it made a mistake 10% of the time. See Axelrod (1984, pp. 36–38).

opponent are in good standing or neither of you is. Then your opponent will cooperate in i and thereafter repeat your last move (play TFT); (2) your opponent is in good standing but you are not. Your opponent will defect in i and thereafter repeat your last move; (3) you are in good standing but your opponent is not. He will cooperate in i and cooperate in $i + 1$ and thereafter repeat your last move.

Situation (1) is the case in round 1. Here you should play TFT. We have seen this before. The proof is the same as the proof given above. In situation (2) if you cooperate in i you will be in situation (1) in $i + 1$. If you defect you will still be in (2) in $i + 1$. So if it is the best move to defect in i, it must be so in $i + 1$, $i + 2$, ... Two sequences then are possible: cooperate, cooperate, cooperate, ..., or defect, defect, defect, ... Since the former gives higher utility (this is implied in the condition of possible mutual cooperation $p > {}^1/_3$), your best choice in (2), as in (1), is to cooperate. Finally, in (3) you are free to defect one round since in $i + 1$ you are again in (1). And in (1) your best choice was to cooperate. This shows that T_1 is in equilibrium. It also shows that it satisfies Lewis' condition (4), for it is the *unique* best reply against itself once there is a small probability of making mistakes in the execution of the strategy.[33] However, it is not the only possible convention. D is now a possible convention as well in this model. If you know the other player will continue to defect regardless of his standing, your unique best reply is to continue to defect as well.

T_1 is a convention. It satisfies all elements of the modified definition. There are more possible conventions besides T_1 and D. T_2, T_3, T_4, ..., T_r, ... all can be stable.[34] T_1, when adopted in this population, satisfies the central requirements of Lewis' definition. First, if T_1 is adopted, requirement (1) *almost everyone conforms to* T_1, is met—be it by stipulation. Note that because of the possibility of mistakes, even the 'almost' clause is satisfied. Second, (2) *almost everyone expects almost everyone to conform to* T_1, is satisfied too, since T_1 is an equilibrium. Third, (3) *almost everyone has approximately the same preferences regarding all possible combinations of actions*, is of course not satisfied. However, that is no problem, since it was dropped from the definition of a convention. Fourth, (4) *almost everyone prefers that almost everyone conform to* T_1 *on the condition that almost everyone conform to* T_1, is satisfied as well, since T_1 is an equilibrium for (almost) all agents in the population. Finally, (5) *almost everyone would prefer that any one more conform to* T_n, *on condition that almost everyone conform to* T_n, is satisfied since we saw that T_1 is not the only possible stable equilibrium.

We can conclude that even in (repeated) situations of considerable conflict of interest, where there are strong incentives for cheating, Lewis conventions could exist. The rule of good standing is an, admittedly crude, example of norms that regulate interactions in such situations have a conventional character. Other game theoretic situations have been analyzed along these lines as well.[35]

5 Intermezzo: Conventions and Moral Norms

Let us pause for a moment and see where the argument has brought us. Are we able to explain some features of moral norms along these lines? On the modified definition of convention, a convention is a stable pattern of interdependent expectations of behavior. The existence of such a pattern is a rule of conduct, a norm, in the group in question. For example, if we apply this to the case of the farmers, T_1 is the rule that says that a farmer ought to provide assistance, provided the other does so as well. One could argue that this rule is also a moral rule as most systems of morality have such norms of mutual assistance. Alternatively, one could interpret T_1 as the rule that promises of mutual assistance are binding. If A announces to B that he promises to help B with his harvest on the condition that she help him, he is announcing that he is following T_1. As we saw above, the unique best reply to T_1 is to follow T_1 as well, so we can see how mutual promises of mutual assistance are really binding. Therefore, the content of a convention can be identical to the content of a moral norm, as the convention of mutual assistance in the farmer example suggests.

What is more, the analysis enables us to make sense of an aspect of moral motivation. Moral agents, it is said, comply with their duty, because it is their duty.[36] They don't merely conform to their duty and act in accordance with the norm. In moral agents, the norm itself is the reason for acting. Note that the analysis of T_1 in the previous section can explain how this could be rational. T_1 is such that compliance (and not just conformity) with it is rational. My reason to comply with T_1 is that others expect me to do so as well. In other words, the reason to comply with T_1 is that it is the established conventional rule. Conventional rules are action-guiding just like moral

[33] In the language of game theory T1 is a *trembling hand perfect equilibrium* in the repeated prisoners' dilemma and TFT is not.

[34] However, not all strategies of the class T are stable. If r is very large, it may pay to switch to play D against such a strategy, depending on the chance of future interactions and the values of Table 2. See Sugden (1986, p. 115).

[35] E.g., Sugden (1986); Binmore (1993); Binmore (1994); Skyrms (1996); Binmore (1998); Den Hartogh (1998, 2002); Verbeek (2002b); Kuhn (2004); Skyrms (2004).

[36] This is a familiar claim in many traditions of ethical theorizing as diverse as that of Kant, Aristotle and Hume.

norms purport to be. This is one of the major attractions of the conventionalist analysis.[37]

However, it is not sufficient to show that a conventional rule can have the same content as a moral norm and be action-guiding in much the same way. Moral norms have additional features which show up when agents deviate from the norm. If a rule is a moral norm, deviance usually is met with criticism, with resentment and indignation. Lewis argued that all conventions become social norms because they are socially enforced because of these forms of criticism.[38] Deviation from the convention will provoke a negative response from the other group members, because they preferred the deviant to act differently and will think badly of him, especially if the deviance is not the result of excusable ignorance or duress.[39] Therefore, this third element, the fact that deviation is met with criticism, can also be explained within a conventionalist theory.

However, it seems crucial for moral norms that this negative response is regarded as justified by those concerned. One could argue that the negative response itself is required by another convention, which gives agents reasons to utter criticism and the like. However, that sets up a regress of interlinking conventions. Each conventional rule has its own conventional rule that requires criticism when it is broken. And of course this second conventional rule has its own rule about what to do if criticism is not forthcoming, etc., etc. It is unclear whether such a regress is vicious and it would certainly be compatible within a broadly conventionalist theory.

Alternatively, one could look elsewhere to justify this negative response. Elsewhere, I have argued that the resentment and indignation with which deviations from a norm are met as well as the guilt that moral agents experience when they violate a norm (without an excuse) presupposes the existence of such virtues as trustworthiness and fairness.[40] The introduction of such virtues is not incompatible with the idea that some moral norms are conventions. In fact, it reinforces the point, since trustworthiness and fairness are responses to expectation about one's behavior. A trustworthy agent will act as she is expected to by the person who puts his trust in her. A fair agent will not let down others who rely on her to do as they expect.[41] These expectations are the result of the established norms and conventions. Such virtues, then, are not *ad hoc* assumptions in a conventionalist theory of moral norms.

We can conclude that the conventionalist analysis has many interesting things to say about moral norms. This makes the objections against conventionalism all the more important. In the next section, I return to them.

6 Are Lewis Conventions Contingent?

Given all this it is tempting to suppose that at least some moral norms are conventional in Lewis' sense. The fact that Lewis explicitly restricted his analysis to games of (pure) coordination is not a reason to suppose that only linguistic rules are Lewis conventions. However, this has not been the most important criticism. The most fundamental objection to the conventionalist project in moral philosophy is that it renders moral norms completely contingent.

Consider the repeated prisoner's dilemma I analyzed above. There, T_1, T_2, T_3, ..., etc., are all possible conventions. So if moral norms are conventional in Lewis' sense, it turns out that other moral norms applying to the same situation are possible as well. It seems a purely contingent fact that T_1 emerged. However, moral norms, it is generally believed, are not contingent in this manner. Consider a fundamental norm, like the one that prohibits murder. It is not by accident that many (if not all) existing moral codes have such a norm. Secondly, given the many possible conventions, Lewisian conventionalism about moral norms seems to imply moral relativism. Which conventional moral norm emerges seems to be bound to contingent factors of the environment and population in which the convention developed. That is not how we think of our moral norms. Moral norms are necessary in some fundamental sense. The same point can be made in less abstract terms. What could be the alternative convention to the norm that forbids murder? How could murder be anything but wrong? Nevertheless, this norm would be contingent if

[37] This also goes some way to show that game theory does have interesting things to say about moral motivation. Often rational choice theorists will claim that moral considerations need to be treated either as inputs in the preference structure of the agents, or as available strategies in the game. The conventionalist analysis shows is that there is another way to account for moral motivation in game theory, namely by focusing on the nature of the reasons of agents to adopt the stable equilibrium strategy. In the example of T1 it is not because the agents have moral preferences that they comply to the norm of mutual assistance, nor because there is a 'moral strategy' in their strategy space. Moral reasons, the analysis T1 of suggests, could be treated as *interdependent* reasons for action. See also Den Hartogh (2002) and Verbeek (2002a, 2007).

[38] Lewis (1969, pp. 99–100).

[39] Similar arguments are made by Sugden (1986, pp. 159–161). Lewis also argued that all established conventions become social norms in this way, but that is an exaggeration. It may be a conventional rule to warn people for the poisonous contents of a bottle with chemicals with a sticker with a skull and bones on a yellow background, but if I write in clear luminescent characters POISON I need not expect a negative response from others.

[40] In Verbeek (2002b, ch. 2 and 4). See also Den Hartogh (2002) and Michael Bacharach (2006, ch. 1).

[41] Verbeek (2002b, ch. 4, 2007).

it were a convention. Therefore, moral norms cannot be analyzed as Lewis conventions.

The force of this objection should not be exaggerated. First, note that on conventional analysis it is not the case that 'anything goes'. Some outcomes could never be achieved (e.g., an outcome where I always cheat and you always cooperate in a repeated prisoners' dilemma). Lewis' analysis sets clear boundaries on what could possibly be a convention. The contingency is not radical. Secondly, it is not true that all conventions are completely contingent. Coordination problems can have superior and inferior equilibria. It is conventional whether we call *Felis catus* a 'cat' or 'un gatto', but a 26 syllable name for this creature is an inferior naming convention. If we loose each other in New York City, we could meet at Grand Central Station or on Times Square, but a convention to meet somewhere in New Jersey is an inferior convention. Therefore, in many cases the possible range of conventions is further delimited by such considerations of superiority.[42] Third, it could be argued that even such fundamental norms as the one prohibiting murder contain conventional elements. Murder is the intentional killing of innocent people. What counts as intentional killing? Who are to be included among the innocent? Is this norm without exceptions? Are all killings of innocent people prohibited (for example, think of cases of voluntary euthanasia)? Does the norm apply equally to all members of a community, or does it allow room for certain people to intentionally kill an innocent person? It is said that in some societies rulers were not prohibited to kill innocent people, while private citizens were not exempted from the prohibition. Furthermore, since these elements vary, it seems that the universality of the prohibition of murder is at best abstract and formal—not substantial.[43]

However, the defender of the idea that moral norms are Lewis conventions has a more fundamental reply to this line of criticism. Remember that for Lewis a convention can emerge when a course of action is salient and this salience is object of common knowledge. Lewis, like Schelling before him, is broad-minded as to what can make a strategy salient. He discusses psychological factors, precedence and explicit agreement as examples of how a strategy can become salient. However, it seems not necessary to stop here. An outcome can become salient as a result of moral reasons other than the norm that is supposed to emerge like a convention. For example, the outcome where parties mutually refrain from murder is morally salient because it has the unique feature that no harm is inflicted, or because it is the unique outcome where the sanctity of life is respected. If we include moral reasons in the factors that can contribute to the salience of an outcome, the resulting convention will not be contingent in an objectionable manner. For this reply to work it is necessary that the norm that this is required is not itself among the moral reasons that contribute to the salience of an outcome. Otherwise, the conclusion that the norm is a Lewis convention does nothing to explain the nature of the norm in question. So it seems that the defender of Lewis has a strong response to the objection that such a conventionalist analysis renders moral norms objectionably contingent.

7 A Dilemma for Lewisian Conventionalists

This reply, however, is a Trojan horse. As I will argue below, it gives rise to a dilemma for Lewis' analysis of conventions. The most obvious way to escape this dilemma brings back the criticism of contingency in full force.

Consider an agent A who has to make a choice between 'top' (T) and 'bottom' (B) in the following simple pure coordination problem.

Table 3 2 by 2 Coordination game

		B	
		L	R
A	T	1, 1	0, 0
	B	0, 0	1, 1

Suppose that 'top left' (TL) is salient (for example, because TL is morally salient). On Lewis theory A is rationally justified in choosing T. Why? T is salient. That is of itself not enough reason to choose T, though it is part of the reason for T. Suppose that A believes that B believes that A finds T salient, then, in addition to the salience of T, there is a further reason for A to choose T. And if A has other higher-order beliefs, he could infer another, further reason to choose T. As a result, on Lewis' account, coordinating on TL is rational and the justification as to why this is the case rests on the idea that higher-order expectations can inform lower-order expectations:

> So if I somehow happen to have an nth-order expectation about action in this two person coordination problem, I may work outward through the nested replications to lower- and lower-order expectations about action. Provided I go on long enough, and provided all the needed higher-order expectations about preferences and rationality are available, I

[42] Thanks to one of the referees of *Topoi* who reminded me of this. However, it is not necessary that only superior coordination equilibria will be selected. See Sugden (1986, ch. 4).

[43] Similar points can be raised about non-moral conventional rules. For example, it is not done to be rude in most, if not all societies. However, by reminding me of this conventional prohibition, you have not told me from which actions I should refrain.

> eventually come out with a first-order expectation about your action—which is what I need in order to know how I should act. (Lewis 1969, p. 31)

So far so good; let's go through the steps of Lewis' account again. First, consider the salience of TL. It cannot be reason to choose T at all. For if it were Table 3 would be an incorrect representation of the predicament of A. If salience would be *a reason* to opt for T, A and B are not immersed in a coordination problem in the first place. For A would have an independent reason to choose T and B, if he believes this to be the case has overwhelming reason to choose L. Thus it looks like the salience of T is not a reason to choose T *at all*. And this is how it should be, since neither A nor B have independent reasons to opt for T or L respectively, given that they are facing a coordination problem.[44]

Secondly, on Lewis' account higher-order reasons can add to the justificatory weight of lower-order (in particular, first-order) reasons. Exactly how do they do this? Well, according to Lewis, the reasoning goes something like this[45]:

(i) A will choose T, on condition that B chooses L.
(ii) A believes that B will choose L.
(iii) Therefore, A has a reason to choose T.

Assumption (i) is not remarkable—it follows from the description of the coordination problem and the assumption that A is rational. Assumption (ii) is more problematic. What justifies (ii)? What is the warrant for A's belief about B's action? It is a second-order belief about B's beliefs about A's actions:

(iv) A believes that B will choose L on condition that A choose T.
(v) A believes that B believes that A will choose T.
(vi) Therefore, A has reason to believe that B has reason to choose L.
(vii) Since A believes that B is rational to a certain degree, A has reason to believe that B will in fact choose L.
(viii) Therefore, A believes that B will choose L, which is assumption (ii).

Again, steps (iv) and (vii) are not remarkable. The assumption that is doing the justifying work is (v), a second-order belief. We can give a similar justification for (v) with non-remarkable assumptions and the third-order belief that A believes that B believes that A believes that will choose L.

At this point, we should get a little worried about the whole account for two reasons. First, higher-order beliefs can only justify lower-order beliefs if these higher-order beliefs have independent justificatory weight of their own. Just like I am entitled to infer q from p and $p \rightarrow q$ if I have reason to accept the premise $p \rightarrow q$ independently of the conclusion that q, I am entitled to infer an nth-order belief from a belief of the order $n + 1$, if there is independent ground for accepting $n + 1$.

Now, *usually* the order of justification of beliefs is bottom up. That is, I am entitled to infer that I believe that I believe that p, if I believe that p; just like my belief that p, is justified if p.[46] So what grounds these higher-order beliefs? Notice that this sets up a regress. This leads us into the second reason for worry: if the order of justification is the usual one mentioned just now, the belief of the order $n + 1$ does not have any further justificatory weight than the belief of the order n. My belief that p is not further justified by my belief that I believe that p. The justification for both beliefs, in the end, is the same, namely, that p.

So it looks like Lewis needs to cut off the regress and do this in such a way that each additional order of belief indeed carries independent extra justificatory weight. At this point Lewis introduces the idea of common knowledge. The idea is that common knowledge of the salience of TL justifies the generation of the higher-order beliefs. To be precise, the salience of TL is common knowledge if and only if:[47]

(i) Both A and B have reason to believe that TL is salient.
(ii) This salience indicates to A and B that each has reason to believe that TL is salient.
(iii) The salience of TL indicates to both A and B, that they will choose T and L respectively.

These three assumptions combined make the salience of TL common knowledge to A and B. With non-remarkable assumptions about the rationality of A and B, each higher-order belief can be inferred

(iv) Therefore, A has reason to believe that B will choose L
(v) And, A has reason to believe that B believes that A believes that B will choose L (and *vice versa*)
(vi) Etc., etc…

In other words, A has ground to accept an n^{th}-order belief about B choosing L, if it is common knowledge that TL is salient. This is why the salience of TL is part of the reason for A to choose T. If TL were not salient, A and B could not have common knowledge of it and the requisite higher-order beliefs could not be justified. This means that the order of justification of the higher-order beliefs is the usual one, from lower to the higher order. Common knowledge of salience could only have this justifying effect if (iii)

[44] See above for an explanation of this sense of independence. See also Den Hartogh (2002, p. 6).

[45] See Lewis (1969, pp. 28–31). I have substituted his formulations with my example where appropriate.

[46] In more formal notation: $B(p) \vdash B(B(p))$ and $p \vdash B(p)$.

[47] Lewis (1969, p. 52).

holds, the salience of TL indicates to both A and B, that they will choose T and L respectively. That is, the salience of TL is already a reason to choose T and L respectively—be it not a sufficient one.

This then generates the dilemma for Lewis. Either salience is a reason for choosing TL, or it is not. (i) If it is, then the higher-order beliefs can be justified under the assumption of common knowledge of the moral salience of TL. However, then we would have to doubt that there is a coordination problem in the first place. The moral salience of TL would already be a reason for A to choose T and a Lewis convention to go to TL would never arise—it would be the only possible point of coordination. (ii) If moral salience is not a reason for choosing TL, then the lower-order beliefs (up to the order 1) could be inferred from the higher-order beliefs and successful coordination will result. However, these higher-order beliefs are not themselves justified. They seem a contingent feature of the agents involved, which brings back the contingency complaint in full force.

8 Implications of the Dilemma

Returning to the question as to whether moral norms can be analyzed as Lewis conventions, we now see that arguing that a convention not to commit murder is the result of the moral salience of such an outcome, gives away the game. For if it is correct, if moral reasons somehow determine the salience of this outcome, the norm not to commit murder could not be a Lewis convention. It would impale the defender of Lewisian conventionalism in ethics to the first horn of the dilemma. If, on the other hand, the defender of conventionalism wishes to avoid this, he will have to argue that the convention not to commit murder is not the result of moral salience. He will then have to accept that this moral norm is contingent. We could have had other moral norms about murder.

So where does this leave the conventionalist approach to moral philosophy? It seems to me that there are at least three responses open to the conventionalist. First, one could bite the bullet and argue that indeed moral relativism is true and that moral norms are contingent. This would embrace the second horn of the dilemma I sketched above. This need not be wrong in the end, but I believe it concedes too much too quickly to the critics of the conventionalist approach. The second line of inquiry that is open to the conventionalist is the one taken by evolutionary game theorists' efforts to analyze morality.[48] In evolutionary game theory individual rationality of the agents is not assumed. All that is assumed is that there is some way in which agents copy successful strategies. Evolutionary game theory can show that in some circumstances, behavioural patterns will evolve which are stable and conventional in the sense that alternative stable patterns are possible as well. The question of rational justification of either the convention or individual compliance does not come up at all in this approach.

Both these versions of conventionalism have in common that they accept that there is no ultimate justification for the existence of the convention. At best, there is a causal story to tell how the convention emerged. In addition, they can give an answer to the question why conventions can and sometimes do persist. Such causal stories, though important, are of limited interest to moral philosophers who want to evaluate entire practices and moral norms.

9 The Legacy of Lewis

This does not mean that Lewisian conventionalism is irrelevant for moral philosophy—far from it. There are many issues and problems in moral philosophy that can be fruitfully tackled with conventionalist models. I want to end this essay by mentioning one, which is an example of the third possible response of the conventionalist.[49] This third response makes a distinction between questions about the justification of convention and questions about the justification of complying with the convention. This response than amounts to arguing that higher-order beliefs have justificatory weight, but denies that common knowledge of salience justifies the existing pattern of mutually interdependent higher-order beliefs that is prevalent among those complying with the convention. The thought is that Lewisian conventionalism cannot really answer the former type of question, but that it has a persuasive answer to the latter type of argument.

Moral philosophers are, among other things, interested in justifying to individual agents why they should comply (as opposed to merely conform) with moral norms. That is, they are interested in analyzing the *authority* of moral norms. Lewisian conventionalism can account for this authority. It can show that individual agents have good reasons to comply with existing moral conventions.

[48] I am thinking of such authors as Sugden (1986); Binmore (1994); Skyrms (1996); Binmore (1998);Vanderschraaf (1999); Kuhn (2004); Skyrms (2004).

[49] This is by no means intended as an exhaustive list of responses. There are many more. For example, Postema in this volume suggests an altogether different interpretation of salience. I have left out the literature on collective intentionality from the discussion since that would complicate the discussion of Lewis enormously. Postema, in this volume, has some interesting things to say on this.

We have already seen an example of how this works in the example of Table 3. Suppose A and B are members of a community that follows the convention "when A choose T, when B choose L". The existence of such a TL-convention means that there exists a pattern of inter-locking mutual expectations that people will follow this convention. This means that both A and B have reason to do their part in the TL-convention. What is more, they are justified in doing so, since both A and B have access to higher-order beliefs about the choice of strategy of the other. In fact the convention, the norm, that prescribes TL is nothing other than a whole set of such higher-order beliefs. These beliefs justify their choosing T and L respectively.

Obviously, these higher-order beliefs are contingent and as a result the convention is to some degree contingent (though the possible range of alternatives to the convention is restricted). However, compliance with such an existing convention is justified in a straightforward manner: it is rationally required. Once it is in place, all members of this community have sufficient reason to comply with it. The fact that this community follows the TL-convention is sufficient reason for each individual member to comply. In other words, the TL-convention has authority over A and B where it comes to their choice of strategy in Table 3. Similar analyses can be given for other contexts of interaction as we have seen. Therefore, in all those instances Lewisian conventionalism can explain and justify the authority of many of our moral norms.[50]

This then suggest how the third response to the dilemma goes. We can accept that higher-order beliefs are free-standing. The way these are generated is essentially causal and not rational. In that sense conventionalism indeed has the implication that moral norms are contingent. However, we need not follow the evolutionary game-theorists in their dismissal of individual rationality. Instead, we can show that rational agents should comply with these conventions even though there may be nothing intrinsically rational about these conventions in the first place.

Lewis' work on conventions continues to inspire moral philosophers to tackle such difficult questions. In the process we sometimes have to modify or even reject elements of Lewis' own theory. However, it is a mark of the importance and continuing legacy of Lewis' *Convention* 40 years after its first appearance, that we still use it as our starting point of theorizing.[51]

[50] This is what I take to be the main contribution of conventionalism in general. See Verbeek (2007).

[51] Many thanks to Govert den Hartogh, Chris Morris, Luca Tummolini as well as the anonymous referees for Topoi.

References

Axelrod R (1981) The emergence of cooperation among egoists. Am Polit Sci Rev 75:306–318
Axelrod R (1984) The evolution of cooperation. Basic Books, New York
Axelrod R (1986) An evolutionary approach to norms. Am Polit Sci Rev 80:1095–1111
Axelrod R, Hamilton WD (1981) The evolution of cooperation Science 211:1390–1396
Bacharach M (2006) Beyond individual choice: teams and frames in game theory. Princeton University Press: Princeton (edited by Gold N and Sugden R)
Bicchieri C (2006) The grammar of society. Cambridge University Press, Cambridge
Binmore K (1993) Bargaining and morality. In: Gauthier D, Sugden R (eds) Rationality, justice and the social contract: themes from "morals by agreement". University of Michigan Press, Ann Arbor, pp 131–156
Binmore K (1994) Playing fair. Game theory and the social contract, vol 1. The MIT Press, Cambridge
Binmore K (1998) Just playing. Game theory and the social contract, vol 2. The MIT Press, Cambridge
Den Hartogh G (1998) A conventionalist theory of obligation. Law Philos 17(4):351–376
Den Hartogh G (2002) Mutual expectations: a conventionalist theory of law. Kluwer Academic Publishers, Dordrecht
Elster J (2007) Explaining social behaviour: more nuts and bolts for the social sciences. Cambridge University Press, Cambridge
Fudenberg D, Tirole J (1992) Game theory. The MIT Press, Cambridge
Gauthier D (1986) Morals by agreement. Clarendon Press, Oxford
Hume D (1998) An enquiry concerning the principles of morals. Oxford University Press, Oxford
Hume D (2001) A treatise of human nature. Oxford University Press, Oxford
Kuhn ST (2004) Reflections on ethics and game theory. Synthese 141(1):1–44
Lewis D (1969) Convention: a philosophical study. Harvard University Press, Cambridge
Mackie J (1977) Ethics. Penguin Books Ltd, London
McClennen EF (1990) Rationality and dynamic choice: foundational explorations. Cambridge University Press, Cambridge
McClennen EF, Shapiro S (1998) Rule-guided behavior. In: Newman P (ed) The new palgrave dictionary of economics and the law, vol 3. MacMillan, London, pp 363–369
Mehta J, Starmer C, Sugden R (1994a) Focal points in pure coordination games: An experimental investigation. Theory Decis 36:163–185
Mehta J, Starmer C, Sugden R (1994b) The nature of salience: an experimental investigation of pure coordination games. Am Econ Rev 84:658–673
Morris CW (1999) What is this thing called 'reputation'? Bus Ethics Q 9(1):87–102
Schelling T (1960) The strategy of conflict. Harvard University Press, Cambridge
Skyrms B (1996) Evolution of the social contract. Cambridge University Press, Cambridge
Skyrms B (2004) The stag hunt and the evolution of social structure. Cambridge University Press, Cambridge
Smith JM (1982) Evolution and the theory of games. Cambridge University Press, Cambridge

Sugden R (1986) The economics of rights, co-operation and welfare. Basil Blackwell, Oxford
Ullmann-Margalit E (1977) The emergence of norms. Oxford University Press, Oxford
Vanderschraaf P (1999) "Game theory, evolution, and justice." Philos Public Aff 28(4):325–358
Verbeek B (2002a) Game theory and moral norms: an overview and an application. Croatian J Phil 2(6):337–352
Verbeek B (2002b) Instrumental rationality and moral philosophy: An essay on the virtues of cooperation. Kluwer Academic Publishers, Dordrecht
Verbeek B (2007) The authority of norms. Am Philos Q 44(3):245–258