

GIACOMO SILLARI

A LOGICAL FRAMEWORK FOR CONVENTION

1. INTRODUCTION

In this paper, I provide a logical framework for defining conventions, elaborating on the game-theoretic model proposed by David Lewis. The philosophical analysis of some of the key concepts in Lewis's model reveals that a modal logic formalization may be a natural one. The paper will develop on the analysis and critique of such concepts as those of common knowledge, indication, and the distinction between epistemic and practical rationality. In particular: (i) the analysis of Lewis's definition of common knowledge reveals that a suitable formalization can be obtained by adopting an approach analogous to that of awareness structures in modal logic; moreover (ii) the analysis of the notion of indication reveals that the agents may be required to make inductive inferences yielding probabilistic beliefs. I shall stress that such aspects, however, pertain to the sphere of epistemic rationality (i.e., they deal with the justification of the agents' beliefs) rather than to the sphere of practical rationality. Confounding the two spheres may lead to the wrong conclusion that, in order to make sense of, say, salience as a coordination device, one should incorporate psychological assumptions into an undivided notion of rationality. On the contrary, practical rationality stands as the usual notion of game-theoretic rationality, whereas epistemic rationality incorporates those aspects pointed out in (i) and (ii) above. This attempt to provide a formal framework for Lewis's theory of convention follows those of Vanderschraaf (1995, 1998) and Cubitt and Sugden (2003). In his work on Lewis, Vanderschraaf provides a characterization of convention as correlated equilibrium, adopting a formal framework close to the set-theoretical one proposed by Aumann (1976). Cubitt and Sugden point out that such a framework does not take into account certain elements that are however present in Lewis's original theory, and propose a different formal setup altogether. In this paper, I show how a formalization based on modal logic can incorporate those distinctive aspects introduced by David Lewis in *Convention*.

The paper is organized as follows: in the following section, I will provide an informal reconstruction of Lewis's account of convention. In Section 3, I will draw the distinction between epistemic and practical rationality and show that Lewis's concept of indication, and his analysis of how common knowledge and higher order expectations come about, pertain to epistemic rationality. In Section 4, I will argue and show that a modal logic formalization of belief, supplemented with awareness structures, can be a natural interpretation of the epistemic concepts involved in Lewis's analysis of convention.

2. LEWIS ON CONVENTION

Lewis's game-theoretic analysis of conventions starts with coordination games. In a pure coordination game, players' interests coincide. As Thomas Schelling¹ put it, games which represent social interactions can be placed along a continuum. One endpoint of such a scale contains games of pure conflict, that is to say, games in which the sum of the payoffs received by the players in every combination of strategies is null (the so called zero-sum games). At the other endpoint are games of pure coordination, in which players receive the same payoff in every strategy combination. Another far more frequent kind of coordination game is one in which players' interests do not exactly coincide, but they still prefer to coordinate with each other.

Lewis's intuition is that coordination problems (non-trivial, in the sense that they have more than one strict Nash equilibrium) underlie every convention, a convention being one particular recurrent equilibrium of such games. Lewis addresses the question of how a specific equilibrium can be reached. In general, in order to have a sufficient reason for choosing a particular action, an agent needs to have a belief (up to a sufficient degree) that the other agent will choose a certain action. Lewis argues that, in the case of coordination games, such sufficient degree of belief is reached by a *system of mutual expectations*. The focus of Lewis's study is on how conventions are sustained, rather than how they originate. In both cases, however, coordination is achieved by means of a system of mutual expectation. What differs is the means by which such systems are produced. There are several ways in which a system of mutual expectations can obtain. A natural one is, for example, that of agreeing to play a certain strategy profile. Another, common coordination device is salience. Facing a new coordination problem, agents

recognize that an equilibrium has certain salient features, and each player expects them to be noticed by the other players, too. A particular kind of salience is precedent; in this case, the salient trait of the equilibrium is that it has served as a solution of a similar coordination problem in the past. Although Lewis does not state the point explicitly, it seems reasonable to conjecture that salience, in general, may serve as a coordination device for originating conventions, whereas precedent is the coordination device involved in their perpetuation.

Since systems of mutual expectations play such a fundamental role in Lewis's theory of convention, it is natural to investigate what mechanisms produce them. Lewis elucidates this point by providing a definition of "common knowledge." Although the term common knowledge² has had remarkable fortune in the literature of so many academic fields,³ it is noteworthy that Lewis is interested in expectations that is to say, beliefs – rather than knowledge. Moreover, his definition does not even immediately deal with beliefs, but rather with reasons to believe, being in fact a definition of "common reason to believe" a certain proposition. However, possibly also due to the fact that the first (and seminal) mathematical formulation of the concept had a natural interpretation in terms of knowledge rather than belief, the expression "common knowledge" became prevalent. Lewis himself later acknowledges the incongruence: "That term [common knowledge] was unfortunate, since there is no assurance that it will be knowledge, or even that it will be true." (Lewis 1978, p. 44, n. 13) I take it that here Lewis is pointing to the fact that reasons to believe may fail to turn into actual beliefs (there is no assurance that it will be *knowledge*), or, even in the case that they do, the agent may entertain false beliefs about the world (there is no assurance that it will be *true*). To preserve conformity with Lewis's original terminology, I shall for now refer to common knowledge as well. This is Lewis's definition:

DEFINITION 2.1. A proposition p is common knowledge in the group G if a state of affairs A obtains such that

1. Everyone in G has reason to believe that A holds.
2. A indicates to everyone in G that everyone in G has reason to believe that A holds.
3. A indicates to everyone in G that p .

Textual evidence shows that the relation of indication should not be interpreted as material implication, but should allow for some kind of inductive inference: in general, if A indicates x to i , that means that, if i had reason to believe that A holds, i would thereby have reason to believe that x holds as well. Clauses (1)–(3), along with suitable assumptions about the agent's reasoning capabilities and inductive standards, originate an infinite series of epistemic propositions such that everyone in G has reason to believe that p , everyone has reason to believe that everyone has reason to believe that p , and so on. The state of affairs A , which allows the agents in the group G to have common knowledge of the proposition p , is said to be a *basis* for common knowledge of p in G .

Lewis's definition of convention is then the following (cf. Lewis 1969, p. 58):

DEFINITION 2.2. 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 any instance of S among members of P ,

1. everyone conforms to R ;
2. everyone expects everyone else to conform to R ;
3. everyone prefers to conform to R on condition that the others do, since S is a coordination problem and uniform conformity to R is a coordination equilibrium in S .

Common knowledge, according to Lewis, should be included in the definition for two reasons: on one hand, on purely descriptive grounds, since it seems that common knowledge is a relevant characteristic of conventions;⁴ on the other, because it prevents certain odd situations to count as conventions as they would in the absence of the common knowledge requirement. For example, if the agents (i) conform to R , (ii) expect everyone else to do so, (iii) prefer everyone else to do so, but at the same time they believe that no one conforms to R *because* they expect others to do the same, then conditions (1)–(3) would be satisfied, but they would not be common knowledge (nor would they, in fact, be first order knowledge). Notice that both such motives to incorporate common knowledge in the definition have nothing to do with the game-theoretic underpinnings of the definition itself.

3. PRACTICAL AND EPISTEMIC RATIONALITY

When trying to coordinate with someone else, one is presented with two different problems. The first is that of forming beliefs about what the other agent will do in the coordination game. The other is to choose an action, based on such beliefs. Hence, in his attempt to provide a rational reconstruction of conventions, Lewis confronts two different issues: one concerns the formation of (rational) beliefs, while the other concerns the choice of (rational) action. That is, certain processes of reasoning of the agents are directed at justifying the choice of a particular course of action, whereas others are directed toward the justification of agents' beliefs. Following Bicchieri (1993, p. 11), we should therefore consider both aspects of *practical* and *epistemic* rationality, the former being relative to the choice of the optimal action with respect to the agent's beliefs and preferences, the latter to the formation and justification of the agent's beliefs in light of available evidence.

It is extremely important to keep the two problems separate, or else the rational reconstruction becomes vulnerable to defeating counter-examples. For instance, substantial parts of Margaret Gilbert's critique⁵ of Lewis's account of convention exploit the lack of a precise distinction of the two kinds of rationality. Consider for example her presumption that "it is natural to (take Lewis to be) assuming the usual game-theoretical approach (to rationality)". (cf. Gilbert 1989, p. 324). The characteristics of such an approach, as they are listed by Gilbert (cf. *ib.*, pp. 321–322) are (i) that agents are perfect reasoners (they use all the relevant information in their possession and make no mistaken inferences), (ii) that they act as reason dictates, and (iii) that they act according to their preferences. Gilbert claims that if rationality is characterized as above, precedent and rationality together are not sufficient to model the behavioral regularities at the core of Lewis's idea of convention. According to her, common knowledge of both rationality and a successful precedent yield no reason to conform to such precedent in the future. If both agents know that there is a successful precedent, they both have a reason to act in accordance to precedent, given that the other will do so, which is the case only if each one knows that the other knows that the other will act in accordance to precedent, and so on ad infinitum. Thus, the infinite regress prevents the players to come up with a conclusive reason for action, unless, Gilbert claims, we incorporate a psychological (and hence foreign to rationality) element into

the picture. Such psychological element would then be an a-rational tendency to follow precedent.

Lewis's analysis, to be sure, tends to keep the distinction between formation of belief and choice of action blurred. For example, cf. the following quote: "The more orders of expectation about action contribute to an agents decision, the more independent justifications the agents will have; and insofar as he is aware of those justifications, the more firmly his choice will be determined" (Lewis 1969, p. 33). This passage can be interpreted as consistent with Gilbert's criticism, which leads to the conclusion that the order of expectations needed to conclusively determine agents' choices is infinite. Lewis, however, does not at any point claim that mutual expectations of conformity to a successful precedent directly constitute a reason to act. He is, rather, saying that they give the players expectations about what other players will do (conform to the precedent), that is to say they provide a reason to believe (since expectation is a particular kind of belief) rather than a reason to act.⁶

In *Convention*, David Lewis does not explicitly characterize the features of epistemic as opposed to those of practical rationality. However, the distinction is crucial in order to avoid criticism *à la* Gilbert. Whereas the characteristics of practical rationality (optimality of action with respect to beliefs and preferences) can be seen as the usual game-theoretic notion of unqualified rationality, the characteristics of epistemic rationality should be more clearly spelled out. I believe that Lewis's text provides numerous insights about the nature of epistemic rationality – even beyond its relevance for explaining the phenomenon of convention – which deserve formal clarification. In particular, two notions elaborated in *Convention* concern the agents' epistemic rationality, that of *indication* and the distinction Lewis is keen to make between *reasons to believe* and *actual belief*. As recalled above, Lewis structures the relation of indication (i) by defining it in terms of reasons to believe and (ii) by implicitly differentiating it from material implication: a state of affairs A indicates proposition *x* to agent *i* if and only if, *if i* had reason to believe A, *i* would *thereby* have reason to believe *x*. The use of the expression *if ... thereby* denotes that Lewis is not thinking of material implication. Moreover, he states clearly that the relation of indication depends on the agents' inductive standards.

The feature of epistemic rationality I am going to focus on in this article is the one that Lewis introduces in his definition of common knowledge via the distinction between reasons to believe and

actual beliefs. In the definition, Lewis gives a sufficient set of condition for an infinite number of epistemic clauses about reasons to believe to arise. Ideally, an agent would believe everything she has reason to, but in practice, obvious limitations occur. Lewis's analysis suggests that the expectations generated by the definition are not actual cognitive states, but merely potential ones. Such potential cognitive state, in my reading of Lewis, is a situation in which an agent may acquire a belief that is epistemically acceptable; that is to say, there exists for her a reason to believe that a certain proposition holds. An agent endowed with sufficient *rationality* acquires then an actual belief out of a potential one (cf. (Lewis 1969), p. 55: "Anyone who has reason to believe something will come to believe it, provided he has a sufficient degree of rationality."). Although Lewis does not qualify what kind of rationality an agent should possess in order to entertain actual beliefs out of her reasons to believe, here, of course he is not thinking of expected utility maximization. The aim the following section is to clarify to what, exactly, Lewis is referring when he assumes that the agents are endowed with this specific kind of rationality.

4. THE FORMAL FRAMEWORK

Is it possible to characterize the distinction sketched in the previous section by formalizing the elements pertaining to epistemic rationality? Cubitt and Sugden (2003) provide the syntax of such a formalization, and incorporate in a formal setup certain elements of the Lewisian analysis of common knowledge and convention overlooked by game theorists and economists in subsequent developments of such concepts. Cubitt and Sugden detect the complexity inherent in Lewis's account: "The concern is with those modes of human reasoning, whether deductive or inductive, that can properly be said to justify beliefs or actions". (Cubitt and Sugden 2003, p. 184.) And moreover: "Lewis' analysis is not, strictly speaking, about knowledge; it is about warranted belief. [...] A belief might be justified according to reasonable standards of inductive inference, yet not be true" (ib.). By and large, it is true that when Lewis's analysis, of common knowledge, has been first formulated in set-theoretical terms (Aumann 1976), its complexities have been levelled out, since the formal model proposed there does not allow for an

object of common knowledge to be false. However, relaxing such assumption need not entail rejecting a set-theoretical formulation *tout court*, as Cubitt and Sugden do in their paper.⁷ As I understand it, their rejection of a set-theoretical approach to modelling epistemic agents is based on the following grounds (cf. Cubitt and Sugden 2003, pp. 206 ff.):

- (i) In partitional models, certain properties of the model are informally considered to be common knowledge among the agents. Since Lewis's analysis is concerned with the origin of common knowledge itself, such models cannot be appropriate.
- (ii) In partitional models, we are modelling knowledge rather than belief. If agent *i* knows *x*, then *x* is the case.
- (iii) Representing the indication relation as material implication (that is – in set-theoretical terms – as set inclusion) eliminates those aspects relative to Lewis's concerns about the agents' (inductive) reasoning.

In my view such objections are not conclusive grounds for dismissing epistemic models cast in set-theoretical (or modal logic) terms. As for (iii), I show in the following how indication can be incorporated in a modal logic setting. As for (ii), Cubitt and Sugden have a point that *partitional* models represent knowledge rather than belief, since the objects of agents' belief must be true in the actual world. However, such requirement, if too stringent, can be dropped. In particular, in the system described later in this section, formulas that are believed by the agents need not be true in the world in which they are believed; the agents can be mistaken.⁸ As for (i), it is essential to stress the difference between the *informal* common knowledge of which the specifications of the set-theoretic model are object, and the idea of common knowledge formally defined and captured in the model. The properties of the model are common knowledge only in the informal sense that they are true in all possible states of the model.⁹ Indeed, in the model there may not be an event expressing such properties, hence the fact that there is informal common knowledge of them need not detract from the relevance and validity of Lewis's formal definition of common knowledge.¹⁰ To be sure, David Lewis's model is not cast in set-theoretical terms, and, indeed, his framework is, by and large, informal. Cubitt and Sugden's formal rendition of it (and especially of its characteristics missing in the usual formalizations) provides us with a very expressive model, although a purely syntactical one. It is the aim of this section to

show that the richness of David Lewis's informal model, and of Cubitt and Sugden's rigorous syntax, may receive a natural and suitable semantical interpretation by means of Kripke (that is to say, set-theoretical) structures.

However, the formalization proposed by Cubitt and Sugden enjoys another feature that is not to be found in the usual set-theoretical models:

- (iv) Following Lewis, they make a distinction between states of affairs and propositions, whereas partitional models deal with events only.

What is to be gained by distinguishing between state of affairs [roughly, as Cubitt and Sugden suggest 'states of the world' in the sense of Savage (1954)] and propositions? The gain in generality is only apparent, since it is trivial to translate the state of affairs A into the proposition 'A holds' and vice versa. Why, then, did Lewis introduce the distinction in the first place? It is my opinion that he did so in order to defend the idea that the indication relation is stronger than material conditional. Recall that saying "the state of affairs A indicates p to i " is tantamount to saying "if i has reason to believe that A holds, i thereby has reason to believe that p is the case." Since Lewis wants the indication relation to be stronger than material implication (at least, in his discussion of the definition of common knowledge he assumes that logical implication entails indication), that would entail that any vacuously false formula indicates any proposition to any agent. By requiring that the indicating formula be a state of affairs, Lewis wants to avoid the paradoxes of material implication being carried over to the indication relation. An attenuation of the impact of the paradoxes can be obtained, without recurring to a sorted language, by requiring that material implication entails indication only in those cases in which the agent has reason to believe the antecedent of the implication (cf. axiom B2 below). Hence, since what I believe to be Lewis's concern motivating the introduction of state of affairs can be taken care axiomatically, I opt for simplicity and drop the distinction altogether.

In sum, the logical framework of *Convention*, including the properties of the indication relations, can be captured by modal axioms interpreted in a Kripke semantics. Agent i 's reason to believe a proposition would thus be represented by means of the modal operator R_i . Notice that I do not intend to explicitly represent here

the process of reasoning conducive to i 's reason to believe a certain proposition p , nor do I intend to assert anything about that process. With the expression $R_i\varphi$ we only capture the fact that, somehow, i would be justified in believing φ . In this sense, an agent's access to a reason to believe a certain proposition can be treated as if it were an agent's propositional attitude towards that proposition, and can be given a natural modal interpretation.

Formally, let us define a language $\mathcal{L}_n^{\Rightarrow}$ whose *alphabet* is the typical alphabet of propositional calculus, augmented with n reason-to-believe operators R_1, \dots, R_n , and indication operators $\Rightarrow_i, \dots, \Rightarrow_n$.¹¹ We shall use the basic connectives \wedge and \neg , and adopt the obvious abbreviations for the others. In particular, $\varphi \rightarrow \psi$ stands for $\neg(\varphi \wedge \neg\psi)$. The countably many *atomic propositions* of $\mathcal{L}_n^{\Rightarrow}$ are denoted by the metavariables p, q, r , etc. and they belong to the non-empty set Φ of atomic (primitive) propositions. The rules for the construction of *well-formed formulas* are the following:

- (i) every atomic proposition p is a formula;
- (ii) if φ is a formula, so is $\neg\varphi$;
- (iii) if φ and ψ are formulas, so is $\varphi \wedge \psi$;
- (iv) if φ is a formula, so is $R_i\varphi$;
- (v) if φ and ψ are formulas, so is $R_i\varphi \Rightarrow_i R_i\psi$.

The formulas of the kind $R_i\varphi \Rightarrow_i R_i\psi$ render Lewis's indication relations, and are to be read " φ indicates ψ to agent i ". The following system based on the language $\mathcal{L}_n^{\Rightarrow}$ captures the deductive core logic of Lewis's *Convention*:

- B0 tautologies of propositional calculus,
- B1 $(R_i\varphi \wedge R_i(\varphi \rightarrow \psi)) \rightarrow R_i\psi$,
- B2 $(R_i\varphi \wedge (\varphi \rightarrow \psi)) \rightarrow (R_i\varphi \Rightarrow_i R_i\psi)$,
- B3 $(R_i\varphi \wedge (R_i\varphi \Rightarrow_i R_i\psi)) \rightarrow R_i\psi$,
- B4 $(R_i\varphi \Rightarrow_i R_i\gamma \wedge R_i\gamma \Rightarrow_i R_i\psi) \rightarrow (R_i\varphi \Rightarrow_i R_i\psi)$,
- B5 from φ and $\varphi \rightarrow \psi$, infer ψ ,
- B6 from φ , infer $R_i\varphi$.

It is important to notice that axioms B1–B6 constitute a minimal system that needs to be further enriched. Although we are representing reasons to believe through modalities, we have not yet specified any property of the R_i operators. Moreover, the axioms proposed above capture the relation between the indication operator and the deductive capabilities of the agents *only*. Nothing is said specifically about the fact that the character of the indication operator is not strict of deductive.

Although modal logic has been used extensively to provide formal accounts of agents' propositional attitudes – knowledge, belief, desire, etc. – its use as an instrument of epistemological inquiry has undeservedly not received as much attention.¹² In this paper I consider modelling reasons to believe as modal operators. Consider the axioms above: *B1* states that reasons-to-believe operators are normal, in that if an agent has reason to believe φ and that $\varphi \rightarrow \psi$, then the agent has reason to believe ψ as well. We find intuitively reasonable that an agent's reasons to believe be closed under *modus ponens*. *B2* introduces the indication relation by linking it to material implication: it states that if φ materially implies ψ and agent i has reason to believe φ , then φ indicates ψ to i . The motivation behind this axiom is that we want to tie material implication and indication together, without letting the paradoxes of the former be carried on to the latter. By requiring that $R_i\varphi$ actually be the case among the premises,¹³ we rule out those situations in which an agent would have a contradictory φ vacuously indicate any proposition ψ to her. Axiom *B3* requires that the indication relation be closed under *modus ponens*, whereas *B4* requires that it be closed under substitution: those are deductive rules with which we want the agents in our system to be endowed, not only when they are dealing with the classical logical connectives, but also when they are considering indications and hence reasons to believe. *B5* is *modus ponens* and, finally, *B6* states that agents have reason to believe all logical truths. Again, it makes sense to require that an agent has reason to believe what logic dictates, although it is not the case that the agents in the model will come to *actually* believe all logical truths.

It is reasonable to add positive introspection to the axioms listed above:

$$\text{B7} \quad R_i\varphi \rightarrow R_iR_i\varphi,$$

since an agent that has reason to believe φ should have reason to believe that she has such reason as well. It also seems reasonable to require that the agents entertain consistent beliefs. This is captured by the axiom:

$$\text{B8} \quad R_i\varphi \rightarrow \neg R_i\neg\varphi.$$

Furthermore, to ease readability, we add the definitional axiom *B9*. It introduces the modal operator R_G , which stands for “every agent in the group G has reason to believe that ...”:

$$\text{B9} \quad \bigwedge_{i \in G} R_ix \leftrightarrow R_Gx.$$

Since we are using a standard modal logic, we can provide a natural semantics for our language in terms of Kripke structures. A Kripke structure is an $(n + 2)$ -tuple $M = \langle W, \mathcal{R}_1, \dots, \mathcal{R}_n, \pi \rangle$ such that W is a set of possible worlds, $\mathcal{R}_1, \dots, \mathcal{R}_n$ are n accessibility relations (one for each agent in the system) on $W \times W$, and π is a truth assignment $\pi := W \times \Phi \rightarrow \{\text{true}, \text{false}\}$ which assigns a truth value to each atom belonging to Φ for each possible world in W . The clauses defining the semantical relation of satisfaction will then be the usual, with $p \in \Phi$ and φ, ψ formulas of the language.

$$\begin{aligned}
(M, w) \models p &\text{ iff } \pi(w, p) = \text{true} \\
(M, w) \models \neg\varphi &\text{ iff } (M, w) \not\models \varphi \\
(M, w) \models \varphi \wedge \psi &\text{ iff } (M, w) \models \varphi \text{ and } (M, w) \models \psi \\
(M, w) \models R_i\varphi &\text{ iff, for all } v \text{ such that } (w, v) \in R_i, (M, v) \models \varphi \\
(M, w) \models R_G\varphi &\text{ iff } (M, w) \models R_i\varphi \text{ for all } i \in G \\
(M, w) \models R_i\varphi \Rightarrow_i R_i\psi &\text{ iff } (M, w) \\
&\models R_i\varphi \text{ and } (M, w) \not\models (\varphi \wedge \neg\psi)^{14}
\end{aligned}$$

How can we incorporate common reason to believe in the system? The standard way to do so in finitary systems is by adopting axioms that characterize common knowledge as a fixed point. In particular, introducing a new operator CR_G that stands for “members of G have common reason to believe that...”, common reason to believe is defined by the axiom

$$\text{B10} \quad CR_G\varphi \leftrightarrow R_G(\varphi \wedge CR_G\varphi),$$

while it is regulated by the rule

$$\text{B11} \quad \text{If } \varphi \rightarrow R_G(\psi \wedge \varphi), \text{ then } \varphi \rightarrow CR_G\psi.$$

It is convenient to express the semantic clause for common reason to believe by using the concept of reachability: we say that a world v is G -reachable in k steps from world w iff there is a path of length k from v to w such that the edges between adjacent worlds are labelled by the accessibility relations of members of G . It follows that

$(M, w) \models CR_G\varphi$ iff $(M, v) \models \varphi$ for all worlds v that are G -reachable from w in any number of steps.

For notational convenience, we define the operator \Rightarrow_G as follows: for all $i, j \in G$, $(R_i\varphi \Rightarrow_G R_i\psi) \leftrightarrow (R_j\varphi \Rightarrow_j R_j\psi)$, that is to say, if φ indicates ψ to an agent $i \in G$, then it does so for any other agent

$j \in G$. Such an operator captures the idea that, although the indication relations differ among agents, in some cases inductive standards are shared by groups of agents, as, for example, in those cases in which common reason to believe comes about.

It is now easy to show that, in our system, Lewis's conditions for "common knowledge" give, in fact, rise to an infinite sequence of reasons to believe;

PROPOSITION 4.1. Let the following three conditions hold:

- (a) $R_G\varphi$,
- (b) $R_G\varphi \Rightarrow_G R_G R_G\varphi$,
- (c) $R_G\varphi \Rightarrow_G R_G\psi$.

Then, the agents in G have common reason to believe that ψ .

Proof. We show by induction on the length of the path that if v is G -reachable from the actual world w , then $(M, v) \models \psi$. Let v be G -reachable from w in 1 step. By $B3$ at w all of $R_G\varphi$, $R_G R_G\varphi$, and $R_G\psi$ hold. Hence, at v all of φ , $R_G\varphi$, and, as desired, ψ hold. By induction hypothesis, if u is G -reachable from w in n steps, then all of φ , $R_G\varphi$ and ψ hold at u . However, from (b), (c) and the fact that $R_G\varphi$ holds at u , it follows that all of $R_G\varphi$, $R_G R_G\varphi$, and $R_G\psi$ hold at u , or that ψ hold at every world $u + 1$ which is reachable in one step from u . \square

We have so far considered agents' reason to believe, rather than their actual beliefs. Lewis's analysis of "common knowledge", though centered on reasons to believe, serves the fundamental purpose of explaining how higher-order expectations (that is, actual beliefs) of the agents come about. Indeed, his rationale to introduce the distinction between reasons to believe and actual beliefs seems to be that of answering the question he poses at p. 52 of *Convention*: "And how is the process (of generating higher-order expectations) cut off – as it surely is – so that it produces only expectations of the first few orders?" The infinite chain of reasons to believe, to which the definition of common reason to believe gives rise, makes no harm descriptively, since it represents only potential epistemic states of the agents, rather than their actual reasoning or beliefs. Thus, Lewis introduces a tension between what a reasoner *should* believe (any proposition she has reason to) and what a reasoner *does* believe

(a subset of the propositions she has reason to). An ideal agent, in this sense, would be unboundedly (epistemically) rational.

Such an agent would have no limitation in all three of her computational power, time, and storage capability. Therefore, she would believe every logical truth, and all consequences of the propositions she believed. What about her inductive capabilities? Whatever her inductive standards might be, she would believe any proposition yielded by such standards. In particular, if according to her inductive standards, a certain proposition were a basis for common knowledge, she would believe the whole sequence of beliefs of infinitely increasing order indicated by it. On the other hand, an actual agent, being boundedly rational, believes only a subset of the logical truths and of the consequences of the propositions she believes. Similarly, she believes only a subset of the propositions yielded by her inductive reasoning and, in particular, of the infinite series of propositions implied by a basis of common knowledge, she believes only those of the first few orders.

Only a portion of the potential (or implicit) knowledge an agent has is translated into actual beliefs. An agent might not focus on a proposition she has reason to believe, and therefore fail to entertain an actual belief about it. This may happen for psychological reasons, or because the proposition, though logically valid, is irrelevant. It may happen because the agent lacks the computational power to actually perform the reasoning necessary to deduce or induce it, or the time to perform the computation, etc. In the case of Lewis's definition of common knowledge, he requires that, for an agent to believe a proposition she has reason to, she possesses a "sufficient degree of rationality" (cf. Lewis 1969, pp. 55–56). Let us spell out the details of his idea. Suppose there is a basis for common knowledge of φ between agents i and j . Agent i has, then, reason to believe φ and, if i has a degree of rationality which is sufficient to realize first-order expectations, i actually believes φ . Also, i has reason to believe that j has reason to believe φ and, if i ascribes¹⁵ to j a degree of rationality which is sufficient to realize first-order expectations, i has reason to believe that j actually believes φ . Provided that i has a degree of rationality which is sufficient to realize second-order expectations, i then actually believes that j actually believes φ . And so on, for all orders of belief. It seems that Lewis is suggesting that, if an agent is endowed with a first-order degree of (epistemic) rationality, she will come to believe everything she has reason to, if an agent is endowed with a second-order degree

of rationality, she will come to believe everything that she has reason to believe that she has reason to, and so on. But this cannot be a satisfactory account of epistemic rationality: an agent may be sufficiently epistemically rational to actually entertain the first-order expectation yielded by a common knowledge basis, but it would be descriptively inadequate to claim that she actually translates in first-order beliefs *any* proposition she has reason to believe.

We can tweak Lewis's intuition about degrees of rationality capturing the relationship between reason to believe and actual belief by means of what is known in the literature as *awareness structures*. The idea of awareness structures is mainly used in the Artificial Intelligence community to represent the distinction between implicit and explicit knowledge¹⁶ (or, as we put it above, between potential and actual belief, that is to say, between possessing a reason to believe and actually believing), while in the economics literature, models of unawareness seem useful in order to take into account unforeseen consequences¹⁷. An *awareness set* is associated to each agent and, intuitively, an agent is said to explicitly know a formula φ if she implicitly knows φ , *and* she is aware of φ (that is to say, if φ belongs to that agent's awareness set.) In our setting, the presence of a formula in the awareness set of a particular agent is witness of the fact that the agent is sufficiently rational to actually come to believe that formula, if she has reason to. According to the argument above, each formula that an ideal agent has reason to believe, up to any degree of epistemic nestedness, should be part of her awareness set and thus actually believed by the ideal agent herself. In practice, limitations dictated by physical constraints (but, possibly, also by constraints related to the agent's heuristics) entail that the set of formula actually believed by any agent is a proper subset of the set of formulas that the same agent, ideally, has reason to believe. By adding awareness structures to the model we gain the ability to formally take in account those limitations.

Formally, on the syntactic level we introduce n new modal operators A_i , one for each agent $i = 1, \dots, n$ in the system, in such a way that the well formed formula $A_i\varphi$ has the intended meaning that agent i is aware of formula φ . Furthermore, we introduce n modal operators B_i , one for each agent $i = 1, \dots, n$ in the system, where the well-formed formula $B_i\varphi$ has the intended meaning that i actually believes φ . As for the semantics, we add to the Kripke structure defined above a set of formulas $\mathcal{A}_i(w)$ for each agent i and for each possible world w . The formulas belonging to $\mathcal{A}_i(w)$ represent those

formulas that agent i is aware of at world w . We can then add the following semantical clauses:

$$\begin{aligned} (M, w) \models A_i\varphi & \text{ iff } \varphi \in \mathcal{A}_i(w), \\ (M, w) \models B_i\varphi & \text{ iff } (M, w) \models A_i\varphi \text{ and } (M, w) \models R_i\varphi. \end{aligned}$$

As for the axioms regulating the behavior of the B_i operators, we add a definitory axiom:

$$\text{B12} \quad B_i\varphi \leftrightarrow R_i\varphi \wedge A_i\varphi.$$

B12 states that an agent actually believes a formula if and only if she has reason to believe it, and the formula is part of her awareness set.

Let us now return to Lewis's definition of common knowledge. Suppose that ψ is a basis for common reason between i and j to believe that φ . Suppose furthermore that the agents are rational up to a certain degree (say second-order rational) and that φ indicates to them that the both of them are first-order rational. By definition of common reason to believe, we then have:

$$\begin{aligned} (1R) \quad & R_i\varphi, \\ (2R) \quad & R_j\varphi, \\ (3R) \quad & R_iR_j\varphi, \\ (4R) \quad & R_jR_i\varphi \end{aligned}$$

and so on. We assume that the agents are rational up to second-order rationality. By resorting to awareness structures, we can precisely spell out such rationality assumption. For the first order we have:

$$\begin{aligned} (1A) \quad & A_i\varphi, \\ (2A) \quad & A_j\varphi, \end{aligned}$$

which, along with (1R) and (2R) yield

$$\begin{aligned} (1B) \quad & B_i\varphi, \\ (2B) \quad & B_j\varphi. \end{aligned}$$

Moreover, ψ indicates to the agents that both of them are "first-order rational". The following proposition captures the indication of first-order rationality:

$$\begin{aligned} (a) \quad & R_i\psi \Rightarrow_{\{i,j\}} R_iA_j\varphi, \\ (b) \quad & R_j\psi \Rightarrow_{\{i,j\}} R_jA_i\varphi. \end{aligned}$$

From (a), (b), (3R), (4R), and the fact that ψ is a basis for common reason to believe, it follows that

- (3') $R_i B_j \varphi$,
 (4') $R_j B_i \varphi$.

The assumption of second-order rationality is expressed by

- (3A) $A_i B_j \varphi$,
 (4A) $A_j B_i \varphi$,

which, with (3') and (4') yields

- (3B) $B_i B_j \varphi$,
 (4B) $B_j B_i \varphi$.

Since ψ does not provide an indication of epistemic rationality higher than first-order, no actual beliefs of an order higher than the second can be inferred.

The clarification of Lewis's assumptions about the epistemic rationality of agents above allows us to consider an example of successful conventional coordination in formal terms. Recall how any instance of the coordination game on which a convention is based presents the agents with a problem of equilibrium selection, and how, according to the analysis developed in the previous sections, Lewis claims that such problem is solved for the agents by means of a system of mutual expectations, i.e., by means of what in general is called "common knowledge". I believe that, against the criticism of Gilbert, such an idea is not inconsistent with that of practically (game-theoretically) rational agents. In particular, what Gilbert calls "a-rational tendencies to follow precedent", are here seen as *epistemically rational* mechanisms to infer which action the other agent might choose. "Common knowledge", or more precisely, any φ that functions as a *basis* of common reason to believe that the other agent will choose a certain course of action, will make the agents aware (in the formal sense) of a solution for the equilibrium selection problem. Intuitively, if φ is a basis for common reason to believe in G that ψ , then we require that ψ belongs to the awareness set of each agent i in the group G . Say that φ represents the fact that there is a precedent according to which, in a situation S , all players conform to the regularity R : φ is then a basis for common reason to believe that ψ , where ψ represents the proposition that the agents conform to R . Then ψ is an element of the set A_i

for each $i \in G$ and, since there is common reason to believe that ψ holds, both $R_i\psi$ and $A_i\psi$, hold for all $i \in G$. According to axiom *B12*, every agent then actually believes that ψ is the case and such a fact, along with the fact that agents are (practically) rational, suffices to explain why players succeed in coordinating and perpetuating the convention. If we denote with φ_G^ψ the fact that φ is a basis for common reason to believe in G that ψ , then the feature of epistemic rationality with which we want to endow the agents is captured by the following requirement:

$$(*) \quad \varphi_G^\psi \rightarrow A_i\psi, \quad \text{for all } i \in G.$$

To see how this fits Lewis's definition of convention, consider a coordination problem S and a solution $\psi := \psi_1, \dots, \psi_n$, where ψ_i stands for "agent i does her part in the coordination equilibrium ψ ". Assume that agents know that ψ has worked in the past as a solution of S . They have reason to believe that ψ solved S in the past. If we denote " ψ has solved S in the past" with φ , we have that

$$(i) \quad R_G\varphi.$$

Suppose such knowledge gives the agents reason to believe that they possess such knowledge (φ is public). We then have that

$$(ii) \quad R_G\varphi \Rightarrow_G R_G R_G\varphi.$$

Finally, assume that the agents have reason to believe that the successful precedent has a bearing on the current situation:

$$(iii) \quad R_G\varphi \Rightarrow_G R_G\psi.$$

Hence, φ is a basis for common reason to believe of ψ in G , and, because of (*),

$$(iv) \quad A_i\psi \text{ for all } i \in G.$$

From (i) and (iii), it follows that $R_G\psi$, hence

$$(v) \quad B_G\psi.$$

In particular, for each $i \in G$, it is true that

$$(vi) \quad B_i\psi_1, \dots, \psi_{i-1}, \psi_{i+1}, \dots, \psi_n$$

that is to say, each agent i (actually) believes that every other player will do her part in ψ . Assuming that agents are (practically) rational, (vi) implies that ψ_i obtains for all $i \in G$, or that the coordination equilibrium ψ will be played. Thus, we have that (1) everyone conforms to ψ , (2) everyone expects everyone else to conform to ψ (vi), and (3) everyone prefers everyone else to conform to ψ since (by our assumption) S is a coordination problem, and ψ is a coordination equilibrium for S . (1)–(3), moreover, are common knowledge in G , hence ψ is a tacit convention in G according to Lewis's definition.

5. CONCLUSION

What is to be gained by rendering Lewis's account of convention in a formal framework? On one hand, a rejoinder against possible criticisms for Lewis's theory of convention. On the other, and, more importantly, formalizing Lewis's framework allows us to uncover relevant epistemological issues, as those related to both the indication relations and the distinction between reasons to believe and actual beliefs. The rational reconstruction of the social phenomenon of convention turns into a vantage point for investigating broader epistemological questions.

Focusing on the distinction between reasons to believe and actual beliefs, one could develop formal models of natural reasoning. Agents' heuristics could be formally captured and analyzed in terms of awareness structures. Awareness structures could prove useful for investigating agents' reasoning about other agents' (epistemic) rationality, yielding a richer approach to interactive epistemology. Focusing on the relations of indication, probabilities and inductive reasoning would enter the picture, providing a more precise account of convention and, again, possibly granting insights toward a philosophically relevant logical approach to epistemology. The basic framework displayed in this paper would of course profit from being enriched and complicated. Its most natural development would consist in incorporating dynamic aspects. One possible avenue by which this might be done would be adding temporal and dynamic modalities [cf. for instance van der Hoek and Wooldridge (2003), or Pauly and Wooldridge (2003)], while another possibility might be by exploring in which way agents process information and revise their beliefs [cf. Bonanno (2005)].

ACKNOWLEDGMENTS

The author wishes to thank Cristina Bicchieri, Horacio Arló-Costa, Peter Vanderschraaf and Robert Sugden for comments, corrections and stimulating conversations. A first version of this paper was presented at the Sixth Conferense on Logic and the Foundations of Game and Decision Theory (LOFT6), Leipzig, July 2004.

NOTES

¹ Cf. (Schelling 1960, p. 84): “If the zero-sum game is the limiting case of pure conflict, what is the other extreme? It must be the “pure collaboration” game in which players win or lose together, having identical preferences regarding the outcome.”

² *Common knowledge* is that epistemic state in which all agents know that p , all agents know that all agents know that p , all agents know that all agents know that all agents know that p , and so on ad infinitum. We shall see that, although Lewis characterizes this concept differently, his definition is equivalent to the iterative one just given.

³ It is difficult to overestimate the influence that the introduction of such an idea has exerted in so many different fields, ranging from economics (Geanakoplos 1992) to computer science (Fagin et al. 1995; Meyer and van der Hoek 1995), from logic [besides the propositional results from Fagin et al. (1995) and Meyer and Hoek (1995), cf. also issues of quantification in Wolter (1999) and Sturm et al. (2002), and the proof theoretical analysis of Alberucci and Jaeger (2005)] to linguistics (Clark 1996).

⁴ Cf. (Lewis 1969), p. 59: “common knowledge of the relevant facts seems to be one (important feature common to our examples of conventions)”.

⁵ Gilbert has attacked Lewis’s definition of convention in several articles. Her arguments are summed up in chapter 5 of Gilbert (1989).

⁶ Cf. (Lewis 1969), p. 31: “Provided I go long enough [...] I eventually come out with a first order expectation about your action – which is what I need in order to know how I should act.”

⁷ Cf., among others (Geanakoplos 1989) (in which the possibility and the implications of representing epistemic states of the agents in set-theoretical models without partitions are explored); (Samet 1990) (in which it is shown that Aumann’s “agreement theorem” holds in models weaker than partitional ones, provided that certain additional conditions are met); (Collins 1997) (which shows that Aumann’s “agreement theorem” holds also when common belief rather than knowledge is assumed, provided that the agents do not entertain false beliefs about their own beliefs).

⁸ Cf. also, with regards to this point (Vanderschraaf 1998), p. 362: “[...] Lewis’s account applies to situations in which the agents’ private information structures are not necessarily partitions”.

⁹ Cf. (Aumann 1999), p. 273, in which is argued that “common knowledge” of the agents’ partitions holds only informally, and is by necessity included in the specifications of each possible world (and, in particular, of the true world).

¹⁰ Even if there exists in the model an event X such that X expresses the properties of the model itself, and there is, thus, formal common knowledge of X , it would not be problematic that there may not exist a basis for common knowledge (in the formal sense) of X since, as Cubitt and Sugden themselves acknowledge in Cubitt and Sugden (2003), p. 190, there can be common knowledge of a proposition without there being a common knowledge basis for it.

¹¹ To avoid notational confusion, it is prudent to emphasize that the operators \Rightarrow_i indexed by agents in the system stand for *indication* and not for material implication, which is represented by the symbol \rightarrow .

¹² The work of Vincent Hendricks is, under this respect an exception. (Cf. for instance Hendricks 2003).

¹³ Since it seems to be the case that, in *Convention*, Lewis takes it a state of affair A to *indicate* to the agents that A holds, the requirement that the agents have reason to believe that A holds if the state of affairs A materially implies p is implicit in Lewis’s account. By explicitly requiring it in axiom $B2$, we may dispense with the distinction and still avoid an indication relation vitiated by the paradoxes of material implication.

¹⁴ Notice that the clause for indication is restricted to the deductive aspect of indication only, since in this article non-deductive capabilities of the agents are not taken into account.

¹⁵ Lewis assumes here that the basis for common reason to believe also indicates the amount of rationality enjoyed by the agents.

¹⁶ Cf. Fagin and Halpern (1988) and Halpern (2001).

¹⁷ Cf. Modica and Rustichini (1994, 1999).

REFERENCES

- Alberucci L. and Jaeger G.: 2005, About cut elimination for common knowledge logics. *Annals of Pure and Applied Logic*, **133**(1–3), 73–99.
- Aumann R. J.: 1976, Agreeing to disagree. *Annals of Statistics* (4):1236–1239.
- Aumann R. J.: 1999, Interactive epistemology i: Knowledge. *International Journal of Game Theory* (28), 263–300.
- Bicchieri, C.: 1993, *Rationality and Coordination*. Cambridge University Press, Cambridge.
- Bonanno, G.: 2005, A simple modal logic for belief revision. *Synthese*, this issue.
- Clark, H. H.: 1996, *Using Language*. Cambridge University Press, Cambridge, MA.
- Cubitt R. P. and Sugden R.: 2003, Common knowledge, salience and convention: a reconstruction of David Lewis’ game theory. *Economics and Philosophy* (19), 175–210.
- Fagin R. and Halpern J. Y.: 1988, Belief, awareness, and limited reasoning. *Artificial Intelligence* (34), 39–76.
- Fagin R., Halpern J., Moses Y. and Vardi M.: 1995, *Reasoning about Knowledge*. MIT Press, Cambridge, MA.

- Geanakoplos, J.: 1989, Game theory without partitions, and applications to speculation and consensus. Technical report, Cowles Foundation Discussion Paper No. 914.
- Geanakoplos, J.: 1992, Common knowledge. *Journal of Economic Perspectives* (6), 53–82.
- Gilbert, M.: 1989, *On Social Pacts*, Princeton University Press, Princeton.
- Halpern, J. Y.: 2001, Alternative semantics for unawareness. *Games and Economic Behavior* **37**(2), 321–339.
- Hendricks, V. F.: 2003, Active agents. *Journal of Logic, Language and Information* (12), 469–495.
- Lewis, D.: 1969, *Convention: A Philosophical Study*. Harvard University Press, Cambridge, MA.
- Lewis, D.: 1978, Truth in fiction. *American Philosophical Quarterly* 15, 37–46.
- Modica, S. and Rustichini A.: 1994, Awareness and partitioned information structures. *Theory and Decision* (37), 107–124.
- Modica, S. and Rustichini A.: 1999, Unawareness and partitioned information structures. *Game and Economic Behavior*, **27**(2), 265–298.
- Meyer, J. J.-Ch. and van der Hoek W.: 1995, *Epistemic Logic for AI and Computer Science*. Cambridge University Press, Cambridge, MA.
- Pauly, M. and Wooldridge M. J. W.: 2003, Logic for mechanism design - a manifesto. In *2003 Workshop on Game Theory and Decision Theory in Agent-based Systems (GTDT-2003)*, Melbourne, Australia.
- Samet, D.: 1990, Ignoring ignorance and agreeing to disagree. *Journal of Economic Theory* (52), 190–207.
- Savage, L.: 1954, *The Foundation of Statistics*. Wiley, New York, NY.
- Schelling, T.: 1960, *The Strategy of Conflict*. Harvard University Press, Cambridge, MA.
- Sturm, H. Wolter F., and Zakharyashev M.: 2002, Common knowledge and quantification. *Economic Theory* **19**, 157–186.
- Vanderschraaf, P.: 1995, Convention as correlated equilibrium. *Erkenntnis* (42), 65–87.
- Vanderschraaf, P.: 1998, Knowledge, equilibrium and convention. *Erkenntnis* (49), 337–369.
- van der Hoek W. and Wooldridge M. J. W.: 2003, Cooperation, knowledge, and time: Alternating-time temporal epistemic logic and its applications. *Studia Logica* **75**(1), 125–157.
- Wolter, F.: 1999, First order common knowledge logics. *Studia Logica* **65**(2), 249–271.