Economics and Philosophy, **33** (2017) 43–72 © Cambridge University Press
doi:10.1017/S0266267116000043 First published online 22 April 2016
cambridge.org/eap

INSTITUTIONS, RULE-FOLLOWING AND GAME THEORY

CYRIL HÉDOIN*

Abstract: Most game-theoretic accounts of institutions reduce institutions to behavioural patterns the players are incentivized to implement. An alternative account linking institutions to rule-following behaviour in a game-theoretic framework is developed on the basis of David Lewis's and Ludwig Wittgenstein's respective accounts of conventions and language games. Institutions are formalized as epistemic games where the players share some forms of practical reasoning. An institution is a rule-governed game satisfying three conditions: common understanding, minimal awareness and minimal practical rationality. Common understanding has a strong similarity with Ludwig Wittgenstein's concept of *lebensform* while minimal awareness and minimal practical rationality capture the idea that rule-following is community-based.

Keywords: Institutions, Rule-following, Epistemic game theory, Common Understanding, Wittgenstein

1. INTRODUCTION

Mostly ignored during a significant part of the 20th century, institutions are now recognized as an important object of study by many economists. The increasing interest in institutions is mainly illustrated by the significant rise of works on the nature and functions of institutions based on rational choice theory and more particularly game theory. Actually, most economists entertain the goal to study institutions with the same set of tools they use to study more 'traditional' economic objects. One of the first explicit game-theoretic accounts of institutions in economics

* University of Reims Champagne-Ardenne (France), Economics and Management Research Center REGARDS, 57B rue Pierre Taittinger, 51096 Reims Cedex. Email: cyril.hedoin@univ-reims.fr. URL: https://sites.google.com/site/cyrilhedoin/

is Andrew Schotter's *The Economic Theory of Social Institutions* (Schotter 2008 [1981]). Since then, many authors have developed this strand of research further (e.g. Aoki 2001; Sugden 2005). Avner Greif's (2006) insightful historical and theoretical study of the institutional foundations of economic development may be seen as the culmination of a research programme that started some thirty years ago.

In two recent articles, Smit *et al*. (2011, 2014) develop what they call the 'incentivized action view of institutional reality'. Though they are not fully explicit on this point, their account can be seen as a rationalization of the methodological and theoretical perspectives of the game-theoretic account of institutions in economics. They contrast their approach with John Searle's theory of institutional facts (Searle 1995, 2010) and argue that the latter is inadequate because it posits the irreducibility of institutional reality. By contrast, the incentivized action view is a 'naturalistic' and bottom-up approach that – the authors claim – has the advantage to account for institutions only on the basis of the incentives motivating individuals' actions. On this view, institutions (or 'institutional objects') are objects individuated by behavioural patterns. The incentivized action view is related to the way institutions are accounted for in a game-theoretic framework. In the latter, institutions are more radically viewed as a behavioural pattern emerging on the basis of the players' preferences and beliefs.

Formally, in this 'Standard Account' of institutions, an institution corresponds to a strategy profile (i.e. a set of strategies, one for each player) such that (i) nobody can increase his expected payoff by switching to another strategy and possibly (ii) the players hold correct and consistent beliefs over what others are doing. This game-theoretic view of institutions actually encompasses a great variety of modelling approaches and substantive assumptions, particularly regarding the nature and the degree of rationality the players are endowed with. Indeed, the Standard Account ranges from evolutionary models of institutions with highly myopic agents to models of repeated games which are based on refined solution concepts such as the subgame perfect equilibrium with highly rational agents playing complex conditional strategies. The former have been first developed by evolutionary biologists[1] and generally rely on a simple aggregative dynamic rule, the replicator dynamics. They have been used by philosophers (Skyrms 1996) and economists (Sugden 2005) to account for the emergence of conventions and norms of fairness. The latter are at the basis of sophisticated theoretical and historical accounts of informal and formal institutions sustaining and organizing trade and other kinds of economic exchange.[2] Between these two extremes,

[1] See in particular the pioneering work of Maynard Smith (1982).

[2] The work of Avner Greif (2006) is probably the most significant example of this approach. See also Milgrom *et al*. (1990) and Greif *et al*. (1994).

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

the Standard Account also includes several intermediate approaches adopting a broadly evolutionary stance while endowing agents with a rationality level somewhere between the myopic assumption of evolutionary game theory and replicator dynamic and the high-rationality requirement of the subgame perfect equilibrium. These approaches refer in particular to the expanding number of studies of learning in games (Fudenberg and Levine 1998). Ultimately, in both the Standard Account and the incentivized action view of Smit *et al.*, institutions do not exist as independent social objects. 'Institutions' is simply a name that is given to behavioural patterns that are salient from the theorist's point of view. This salience partially comes from the fact that these patterns can be conceptualized as self-enforcing in a game-theoretic framework. But – and this is the key point – the concept of institutions does not play any distinctive role in the *explanatory* endeavour. As a scientific and explanatory concept referring to some part of the social reality, it can simply be eliminated from the economist's vocabulary.

The Standard Account can be characterized then as 'reductionist' in the sense that it is possible to account for an institution by referring to simpler, more basic, concepts. We could also say that it expresses a 'summary view' of institutions. Indeed, there is an obvious link between the Standard Account and what John Rawls (1955) called the summary view of rules, i.e. rules as summaries of activities. One significant consequence is that institutions are no longer distinctive properties of *human* societies: if institutions are nothing more than behavioural patterns generated by the appropriate incentives then, under a sufficiently loose definition of the concept of incentives, animals and bacteria also have institutions. The fact that game theory is widely used by biologists and other scientists confirms that formally speaking, there is nothing specific to human institutions so defined. As such, this is not really an objection to the Standard Account. However, the specificity of humans regarding their thinking abilities and the way they organize at the social level is well documented (e.g. Tomasello 2014). One may believe that what sets human societies apart is their ability to create institutions. If this is true, then the Standard Account is unable to account for this ability.

My main objective in this paper is to propose an alternative conception of institutions on the basis of a theory of rule-following in games. In a rule-following account, individuals behave as they do *because* of a rule. The point is that many institutions are sustained by what can be called constitutive rules (Rawls 1955; Searle 1995), i.e. rules that define a practice. Many behavioural patterns can then be explained by the fact that the agents have some knowledge of these rules and that, under the appropriate epistemic conditions, this knowledge is sufficient to lead them to behave in a certain way. Among these conditions features the requirement that the agents must be confident in the fact that they

infer the same practical conclusions as others from a given state of affairs. I refer to this condition as the fact that the agents have a *common understanding of the situation* and I show that it is an integral component of any constitutive rule. Moreover, it has strong affinities with Wittgenstein's notion of *lebensform* (Wittgenstein 2010 [1953]).

The article is organized as follows. In the second section, I present in an informal way the notion of rule-following behaviour as it appears in Ludwig Wittgenstein's and David Lewis's writings. The third section develops a general game-theoretic and epistemic framework. The fourth section specifies a set of axioms regarding the players' epistemic states and reasoning modes that define rule-following behaviour. The fifth section defines formally an institution as a rule-governed game and discusses several conditions and issues related to this definition. It investigates in particular the role played by communities in enforcing rule-following behaviour. The sixth section briefly concludes by examining one potential objection to my account. Finally, an appendix features a summary of the main axioms and conditions used in the paper, as well as an example providing a simple illustration of the formal analysis.

2. GAME THEORY AND RULE-FOLLOWING BEHAVIOUR

Following the Standard Account, what game theorists are claiming to account for under the name 'institution' are actually self-enforcing behavioural patterns that are the product of undetermined reasons for action. This paper builds on a postulate regarding this claim: what game theorists call 'institutions' under the Standard Account do not correspond to what many social scientists actually refer to when they speak of institutions. To be more specific, for many social scientists and philosophers of social sciences, institutions are not only things to be explained (i.e. a component of the *explanandum*), they figure also as an *explanation* of many social phenomena and individual behaviour (i.e. a component of the *explanans*). In the latter case, the reductionism of the Standard Account is necessarily wrong-headed because obviously we need the concept of institution as part of the explanatory endeavour. My aim is to show that, while remaining in a game-theoretic framework, a 'thicker' concept of institution can be developed on the basis of a theory of rule-following. Concretely, my purpose is to show how we can account for the fact that institutions, because they provide a *reason for action*, can be explanatory in a game-theoretic framework.[3]

[3] I do not intend to claim that what I call the Standard Account is 'wrong' or irrelevant. Rather, I want to suggest that there are mechanisms in the social world that the view of institutions as behavioural regularities ignores or at least downplays. The comparative relevance of the Standard Account and my rule-following account depends on one's

The notion of rule-following finds its roots in Ludwig Wittgenstein's masterpiece, *Philosophical Investigations* (Wittgenstein 1953). Wittgenstein devoted a significant part of his book – essentially paragraphs ##138–242 – to the development of a series of thoughts about what constitutes the fact of following a rule.[4] Though Wittgenstein was essentially concerned with the nature of languages and their related rules, his account of rule-following has obviously a larger scope of relevance. Indeed, many of his examples are about social activities other than speaking a language. The notion of rule-following as developed by Wittgenstein also has strong connections with the concept of *constitutive rules* first developed by Rawls (1955) and later by Searle (1995). Constitutive rules are rules that define a practice and make it possible. More formally, a rule R is constitutive of a practice P if P-ing consists in following R. If we accept the idea that at least some institutions are grounded on constitutive rules, then it is straightforward that the Standard Account (or Smit *et al.*'s incentivized action view) is unable to account for them. Indeed, in the case of constitutive rules, it is impossible to explain a practice without referring to one or several rules given that the nature of the practice consists in following them. Examples of constitutive rules are many. For instance, hitting a home-run or the more general practice of 'playing baseball' is nonsensical without making a reference to the rules that define baseball. Similarly, one would have a hard time explaining the behaviour of a person throwing a ball at another person trying to hit it without referring to the rules of baseball. Another illustration is the fact of buying something with some pieces of paper: the very act of buying relies on the existence of rules regarding what counts as a 'good' or as 'money'. Moreover, as in the case of baseball, it seems difficult to explain the behaviour of buyers and sellers without taking into account the rules that define a market exchange.[5]

perspective: for instance, the former is probably more useful to explain the *emergence* of institutions. But the latter is required to explain how institutions provide a reason for action.

[4] Bloor (1997) offers one of the most ambitious attempts to derive from Wittgenstein's account of rule-following a theory of institutions. Some loose connections with game-theoretic reasoning and concepts transpire in several places in Bloor's book but the author did not intend to make them explicit. This article can be seen as an attempt to go further in the formalization of these connections.

[5] Elsewhere (Hédoin 2015), I argue that in the Standard Account of institutions, all rules are merely regulative, i.e. they do not determine the nature and the content of a practice but only select one among the many behavioural patterns (equilibria) the practice may lead to. On the account defended here, *all* institutions depend on constitutive rules. Whether or not we can distinguish between constitutive and regulative rules is a difficult question that I cannot address fully here. A growing consensus in the literature is that this distinction is dubious because it depends on the way a given practice is described (Hindriks 2009; Hindriks and Guala 2015). Then, all rules would be both constitutive and regulative. I am

It has been argued by some authors that rational choice theory and in particular game theory are unable to account for all kinds of rule-following behaviour (e.g. Vanberg 2004). Lahno (2007) notes that while the notion of rule-following is not totally alien to rational choice theory, it is related in a very restricted sense. If rule-following behaviour consists in acting on the basis of a practical rule, then the rational agents of rational choice theory are indeed acting on the basis of a practical rule, namely utility maximization. Lahno (2007) calls any practical rule reducible to this maximization principle an 'instrumental rule', i.e. a rule that serves the sole function of utility maximization. However, rule-following behaviour cannot be reduced to instrumental rules, particularly in a game-theoretic framework. First, in many cases, 'tie-breaking rules' are needed to help the agent to make a choice between two or more alternatives that have the same rank in the agent's preference ordering. More significantly, agents must often use 'coordination rules', i.e. rules that permit the agents to solve coordination problems (Lahno 2007: 444). Interestingly, while rational choice theory seems unable to account for this kind of rules, it is also perfectly rational in the very terms of rational choice theory to follow coordination rules (Lahno 2007: 446).

However, these judgements about the possibility of accounting for rule-following behaviour in a game-theoretic framework seem too negative. Starting from Wittgenstein's writings and on the basis of David Lewis's theory of conventions (Lewis 2002 [1969]), Giacomo Sillari (2013) makes a convincing argument that rule-following can be captured as a kind of conventional agreement between preferences and beliefs among the members of a community. Still, because Sillari's account puts a major emphasis on Lewis's claim that conventions mainly arise thanks to the 'force of the precedent' (i.e. the history of plays in the game), it retains a key feature of the Standard Account: if an institution (or a rule) provides agents with a reason for action, it is only through a behavioural pattern to which it is ultimately reducible. However, as he notes himself, an agent may infer from the same behavioural pattern an infinite number of practical conclusions regarding what he should do. I therefore think that Sillari's account is on the right track because it puts an emphasis on the most relevant issue for a theory of rule-following: how agents *reason from* a given state of affairs to infer a practical conclusion about what they should do.[6]

agnostic on this point but if we can indeed distinguish between constitutive and regulative rules, then this reinforces the claim that the Standard Account and my rule-following account are complementary.

[6] At this point, I have to warn the reader regarding the way I interpret the notion of reasoning as well as any other notions pointing to intentional states (preferences, beliefs), particularly because these notions could seem to be in contradiction with Wittgenstein's

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

I contend that this issue can be adequately dealt with in a game-theoretic framework. Like Sillari (2013) but also Cubitt and Sugden (2003), I think David Lewis's writings on conventions provide the right starting point. More particularly, it is Lewis's theory of *common knowledge* that is particularly relevant here. Indeed, Lewis was the first to provide a detailed account of how a proposition can become common knowledge among the members of a population. He uses the key concept of *indication*: a state of affairs A indicates to an agent i that a proposition P holds if, whenever i has reason to believe that A holds, he thereby has reason to believe that P holds (Lewis 2002 [1969]: 52–3). The indication notion is all about practical reasoning and particularly *inductive* reasoning. Accordingly, it should be at the core of any theory of rule-following; indeed, Wittgenstein's writings suggest that *to follow a rule is the same as to behave on the basis of an inductive mode of reasoning that each person knows or has strong reason to believe that it is shared by all the members of a given community*. The notion of rule as used here is a 'thick' one, in the sense that it is not reducible to a behavioural pattern. Basically, a rule points to what should be done. It can do so in many different ways. For instance, Lewis entertained a distinction between 'conventions', 'rules' and 'social norms', pointing out that only the first do not need some kind of enforcement through sanctions or normative pressure. Moreover, according to Lewis, conventions (contrary to rules) rely on some kind of salience (salience from precedent) and on the fact that it is commonly known that they hold. Finally, Lewis clearly identifies conventions with behavioural regularities and nothing more, which makes his notion of convention also compatible with the Standard Account. My use of the notion of rule departs from Lewis's and is related to the fact that conventions, rules (in Lewis's sense) and social norms are all specific kinds of institutions. Whether or not an institution depends on an explicit enforcement mechanism through sanction or normative pressure, or relies on some kind of salience and tacit agreement, or implies some kind of arbitrariness is secondary here. In all the cases, I contend that institutions, including conventions, can be characterized in terms of rule-following behaviour in a game-theoretic and epistemic framework. Thus,

ontology. My understanding of intentional states in this paper is broadly 'externalist' in the sense defended by Ross (2014) on the basis of Dennett's intentional stance functionalism. According to externalism, to say that an agent 'reasons that X' or 'intends to Y' is not to make a particular claim about the state of the agent's brain and the possibility that this agent is in some state of consciousness. Nor is it to claim that the agent should be able to verbalize his reasoning or his intention. Rather, when we say that an agent 'reasons that X' or 'infers B from A', we are taking the intentional stance, i.e. an epistemological posture that helps us to make sense of the agent's behaviour. In this sense, any intentional state is the product of a complex interaction of the agent's behaviour, the institutional context this behaviour is embedded in, and the analyst's scientific goal. See Ross (2009) for the claim that Wittgenstein must be read as an externalist.

my account is not a Lewisian one, though it takes advantage of Lewis's key concepts of indications and reason to believe.

3. RULE-FOLLOWING IN A GAME-THEORETIC AND EPISTEMIC FRAMEWORK

This section and the following one develop a game-theoretic account of rule-following broadly based on Lewis's writings on conventions and common knowledge and on Wittgenstein's views developed in *Philosophical Investigations*. I do so by embedding a standard game-theoretic framework into an explicit epistemic modal logic system. The reasons for proceeding in this way will appear as the section proceeds but can already be stated as a starting point. Though Lewis's and Wittgenstein's accounts largely differ in their details, both see rule-following as a behavioural event that finds its roots in expectations that are constitutive of a practice. More exactly, the persons' behaviour corresponds to a practice that entails the knowledge or the belief that some rule holds and where the rule's meaning is defined by the very practice it is constitutive of. In other words, rule-following has two dimensions. Firstly, a dimension of *reasoning*: to follow a rule is a fact about the way players infer from an event or a state of affairs a practical conclusion regarding what they should do. It is here in particular that Lewis's notion of indication is relevant. Secondly, a dimension of knowledge and/or belief: to follow a rule implies that the player knows the rule and/or believes that the rule holds and hence has a specific knowledge of the other players' mode of reasoning. More precisely, following Lewis, I argue that following a rule depends on the fact that there is a common reason to believe that a rule holds in the relevant population. My main point will be the following: in the case of rule-following, beyond the common reason to believe that some strategy profile is implemented, some kind of 'agreement' over the players' reasoning is also required. When this latter condition holds, I will say that the players have a *common understanding of the situation*. I shall argue that such a common understanding is constitutive of rule-following behaviour and is formally identical to Wittgenstein's concept of *lebensform*.

The above considerations justify to combine game theory with interactive epistemology and more generally with epistemic modal logic to deal with the rule-following phenomenon.[7] This leads naturally to

[7] The encounter of game theory and epistemic modal logic is sometimes referred to as the 'epistemic program' in game theory. The epistemic program can be considered to originate in Aumann's (1976, 1987) articles on common knowledge and correlated equilibria. It has been largely developed since, in particular as a way to formally state the epistemic requirements of various solution concepts in normal and extensive form games. See for instance Perea (2012).

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

the main proposition of the paper, i.e. that it is relevant and insightful to formalize institutions as epistemic games to account for behavioural regularities in a population. The purpose of this formalization is to shade light on the distinctive epistemic mechanisms on which the working of many (if not all) institutions relies. If, as I suggest above, at least some institutions indeed depend on constitutive rules, then I propose that the modelling approach pursued here helps to make clear that these institutions do not reduce to behavioural patterns but correspond to whole practices, including the way people think and reason about a given situation and about what others are doing. By identifying institutions with behavioural regularities, the Standard Account arguably ignores or at least downplays this feature which is common to the working of many institutions.

There are many ways to proceed as well as many kinds of formalization available depending on the axioms retained at the syntactic and/or semantic level. In the following, I will first propose a generic game-theoretic and epistemic framework along the lines suggested by Bacharach's 'broad game' concept (Bacharach 1987, 1994). This framework is very flexible and allows one to formalize the epistemic properties of a game both semantically and syntactically while remaining agnostic regarding the specific axioms retained in the epistemic modal logic. On this basis, I propose then to characterize rule-following behaviour through an epistemic game using the axioms of *KD45* modal logic. Later in the paper, in section 5, I introduce the concept of awareness structures to formalize actual beliefs. As suggested by Sillari (2005), the latter are a convenient way to account for the Lewisian distinction between actual beliefs and reason to believe. The appendix features a summary of the main axioms and principles.

We start from a game G described by the tuple $< N, S, \phi, \{u_i\}_{i \in N} >$. $N = \{1, \ldots, n\}$ is the set of players. $S = S_1 \times \ldots \times S_n$ is the set of pure strategies defined as the Cartesian product of the available pure strategies for each player i. We denote X as the set of physical outcomes. The function ϕ: $S \rightarrow X$ maps any strategy profile $\boldsymbol{s} = (s_1, \ldots, s_n)$ into an outcome $x \in X$. The players' well-ordered preferences are represented by a utility function u_i: $X \rightarrow \Re$ mapping each outcome into some real number. We do not assume that the players are necessarily Bayesian rational. The reason is that we do not want to commit ourselves to a specific kind of practical rationality and this allows for a variety of reasoning modes. The point is to relate the players' behaviour in this game to some rule or institution; more precisely, I shall show how to account for the fact that players behave in a specific way *because of the fact that they have reason to believe that a particular rule holds*. To do this, it is necessary to complete the description of G with a *broad theory* $\mathcal{J}$ of G which itself consists in an ordered pair $\mathcal{J} = (\Theta, \text{NL})$, where Θ is a theory of rational play in G and NL is a set

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

of non-logical axioms for Θ. The resulting tuple $< N, S, \phi, \{u_i\}_{i \in N}, \mathcal{J} >$ describes what I call an epistemic game $\mathcal{G}$. Following Bacharach (1994), I will characterize the broad theory $\mathcal{J}$ as a set of axioms and theorems about what the players in G do, know and believe. These axioms may be usefully expressed on the basis of a formal language, i.e. a syntax.[8] The syntax consists of an alphabet with the following elements (Bonanno 2002): (1) a finite or countable set A of atomic propositions p, q, t corresponding to sentences; (2) the traditional connectives $\neg$ (for 'not') and $\vee$ (for 'or');[9] (3) a set of m modal operators $\Box_1, \ldots, \Box_m$. In epistemic modal logic, the usual modal operators will be k_i and b_i standing for 'i knows' and 'i believes' respectively. A *formula* is a finite string of symbols formed by combining connectives and modal operators from the atomic propositions. We denote F the resulting set of formulae.[10]

The broad theory $\mathcal{J}$ of G in now characterized in the following way. The theory of rational play Θ is defined by a set of formulae or *theorems* $F_\Theta(G)$ obtained on the *minimal* basis of

- A set NL of non-logical axioms and postulates about the players' knowledge or beliefs, and principles of rational choice.
- A set of *practical statements* of the kind $D_i x$ ('i does x'), for all $i = 1, \ldots, n$.
- The connectives $\neg$ and $\vee$.
- A set $\{\Box_1, \ldots, \Box_n\}$ of epistemic operators.
- A set of logical axioms
  - N: $\forall p \in F_\Theta(G)$, $p \rightarrow \Box_i p$, for $i = 1, \ldots, n$.
  - K: $\Box_i p \wedge \Box_i (p \rightarrow q) \rightarrow \Box_i q$.

It is important to acknowledge that these are only minimal principles to characterize the theory of rational play in G. Indeed, in the following, I will add more structure for instance to characterize rational play in G or stronger assumptions regarding the epistemic reasoning abilities of the players. For the moment, note that the two logical axioms N and K are fairly standard in modal logic. The former states that if p is a theorem of Θ, then each player knows/believes/has reason to believe this and that this is also a theorem of Θ. The latter is called the knowledge axiom

[8] The broad theory $\mathcal{J}$ of G is conceptually similar to the concept of 'interactive reasoning scheme' developed by Cubitt and Sugden (2014) to account for Lewisian common knowledge.

[9] The connectives $\wedge$ (for 'and') and $\rightarrow$ (for the material implication 'if … then') are derived from $\vee$ and $\neg$ in the following way: $p \wedge q$ corresponds to $\neg(\neg p \vee \neg q)$ and $p \rightarrow q$ corresponds to $\neg p \vee q$.

[10] By the definition of the notion of formula, the set F is closed under negation, disjunction and the modal operators, i.e. a proposition is a member of F if and only if it is obtained by combining the atomic propositions with $\vee$, $\neg$ and $(\Box_1, \ldots, \Box_n)$.

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

and indicates that each player knows/believes/has reason to believe the logical implications of what he knows.

The set NL of non-logical axioms for Θ provides statements about the players' knowledge and beliefs, as well as their practical reasoning abilities. For instance, it is standard in game theory to assume that each player knows his set S_i of available strategies as well as his reflexive, complete and transitive preferences over outcomes. Another standard assumption is that players are rational. More generally, I will assume throughout this paper that the whole game G is commonly known by all the players in N.[11] The set of formulae derived from these non-logical axioms is denoted $F_{NL}(G)$. What should be remarked here is that a formulae f which belongs to $F_{NL}(G)$ also belongs to $F_\Theta(G)$, i.e. $F_{NL}(G) \subseteq F_\Theta(G)$. Axiom **N** above then implies that whenever $f \in F_{NL}(G)$, then $\Box_i f \in F_\Theta(G)$. $F_\Theta(G)$ thus corresponds to the set of propositions and formulae that are known or believed by the players regarding the way the game G is played.

The above syntax can be given a semantic counterpart (Bonnano 2002). It takes the form of a possible world (or states of the world) structure defined in terms of truth statements: it states at each state of the world whether a given sentence (i.e. propositions and formulae) is true or false. First, we define a state space Ω whose components are states of the world ω. A state ω is a complete description of everything that is relevant for the modeller. Any subset $E \subseteq \Omega$ is called an event. An event holds (or happens) at ω whenever $\omega \in E$. The relation between states is captured by a binary accessibility relation $R_i: \Omega \rightarrow 2^\Omega / \varnothing$ defined for each person i (2^Ω is the set of events) which may satisfy several properties (see below). Therefore, $\omega R_i \omega'$ means that state ω' is accessible from state ω for person i. I denote by $R_i(\omega)$ the set of all states that are accessible for i at ω.[12] The resulting tuple $\{\Omega, R_1, \ldots, R_n\}$ is called a *frame* $\mathcal{F}$. The link between the syntax system and the semantic system is provided by the notion of *model* (Bonanno 2002). For any frame $\mathcal{F}$, a model M provides a transcription into a syntax on the basis of a function $V: A \rightarrow 2^\Omega$. V associates with each atomic proposition $p \in A$ the set of states at which p is true. Then, for any formula f, it is possible to determine whether f is true or false at ω in model M, which we denote $M, \omega \models f$ and $M, \omega \not\models f$ respectively. The relationship between the modal operators $\Box_i$ and the accessibility relations R_i is derived as follows:

$$M, \omega| = \Box_i f \text{ if and only for all } \omega' \text{such that } \omega' \in R_i(\omega'), M, \omega'| = f.$$

Finally, we say that a formula is *valid* in model M if and only if it is true at every state $\omega \in \Omega$. All the valid formulae in M are theorems

[11] More exactly, using the common reason to believe operator r^* of the next section, I will assume that there is common reason to believe that the game G is played.

[12] $R_i(\omega)$ is sometimes also called a possibility set.

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

(or tautologies) of $F_\Theta(G)$. Focusing on the specific case of epistemic modal logic, this implies that we can construct semantic equivalents to the epistemic operators defined in the syntax. For instance, for any syntactic knowledge operator k_i, we may define a semantically equivalent knowledge operator K_i. For any formula f such that $M, \omega \models f$, axiom N implies that player i knows that formula f is true at ω if and only if f is true at every accessible state:

$$k_i f = \{\omega | \forall \omega' \in R_i(\omega) : M, \omega' | = f\}.$$

Now, denote E the event 'f is true', i.e. $M, \omega \models f$ for every $\omega \in$ E. Then, i knows E at ω is stated as

$$K_i(E) = \{\omega | R_i(\omega) \subseteq E\}.$$

Note that K_i(E) is itself an event. More generally, if the syntax is sound and complete with respect to M, all formulae in $F_\Theta(G)$ correspond to events in M. Beyond the fact that it is more convenient to use, the semantic approach has also the virtue of making clear that the broad theory $\mathcal{J} = (\Theta, \text{NL})$ of G actually provides an information structure related to G: it states what each player knows and believes about each other players' behaviour, knowledge, beliefs but also practical rationality and reasoning modes depending on the situation. Therefore, the resulting epistemic game $\mathcal{G}_\mathcal{J} : < N, S, \phi, \{u_i\}_{i \in N}, \mathcal{J} >$ corresponds to the description of what Aumann and Dreze (2008: 72) call a 'game situation': 'a game played in a special context' and where 'a player's expectation depends upon the context – the 'situation''. The notion of game situation has a very strong Wittgensteinian stance because it emphasizes that how one plays in a game depends on a specific situation. From the perspective of Wittgenstein's language game, that means that words or signs do not have a deterministic meaning; meaning depends on what each knows (or believes) about the conventional practice or use in this specific situation.

4. RULE-FOLLOWING, COMMON REASON TO BELIEVE AND INDICATION

I have provided a very general framework that allows for a complete description of how a game is played depending on the players' practical and epistemic rationality standards. The broad theory (or information structure) $\mathcal{J}$ provides a self-contained account of what happens in a given game. Moreover, axioms N and K imply that the players know/believe (or at least have reason to believe) the broad theory. An epistemic game thus provides a complete description of the features that are constitutive of a situation and which are *in principle* accessible to the players themselves.

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

However, to provide a rule-following account of institutions along the lines suggested by Lewis and Wittgenstein, more substantive assumptions are required. Though it is not my aim to give an exegesis of Lewis's and Wittgenstein's writings, specific assumptions regarding the nature of both the epistemic states and the practical rationality standards of individuals have to be made. Clearly, these assumptions significantly depart from the game theorist's usual assumptions.

I will proceed by extending the set of axioms made in the broad theory $\mathcal{J}$ of some game G. A distinction should be entertained between three kinds of axioms: the logical axioms that account for the players' deductive abilities, the epistemic axioms or postulates that account for the epistemic states of the players and the practical axioms that refer to the players' practical rationality and reasoning modes. While the first set of axioms pertain to the theory of rational play Θ, the other two belong to the set NL, even though epistemic postulates are also related to properties of the epistemic operators used. As a first point, as Cubitt and Sugden (2003) repeatedly emphasize, the use of knowledge operators seems unable to capture most of the essence of Lewis's theory of common knowledge and conventions. Knowledge is defined as true belief and is generally captured by the 'truth axiom' $k_i p \rightarrow p$ which states that everything person i knows is true.[13] Combined with axioms N and K, the truth axiom entails what is sometimes called 'logical omniscience', i.e. 'the principle that if something is a theorem, then rational knowers know it' (Bacharach 1994: 14). The standard S5 modal logic also endows players with positive and negative introspection abilities, meaning that one knows that he knows something but also that he knows that he does not know something. This entails logical and deductive abilities that seem to be far beyond the reach of normally rational people. Arguably, this is not a satisfactory way to account for rule-following behaviour, even though the resulting semantic frame is still a useful and convenient way to formalize common knowledge in a population of ideally rational agents.

Following Lewis, a more realistic way to capture players' epistemic states is to reason in terms of 'reason to believe'. Informally, a person i has reason to believe proposition p if provided that he is aware of p, he should actually believe p. This does not mean that person i *actually* believes p because of cognitive limitations or other reasons but under appropriate conditions, reason to believe leads to actual belief. Formally, I denote the sentence 'i has reason to believe' by the syntactic epistemic operator r_i. The N axiom then means that all persons i have reason to believe propositions and formulae in $F_\Theta(G)$ and the K axiom that all persons i have reason to believe the logical implications of what they have reason to believe. Since

[13] The semantic counterpart for the truth axiom is $K_i E \subseteq E$.

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

the truth axiom is clearly inappropriate in the case of reason to believe, we substitute it for the weaker D axiom,

D $r_i p \rightarrow \neg r_i \neg p$

Axiom D states that the players' reasons to believe have to be consistent: if one has reason to believe p, then he cannot have reason to believe not p.[14] We also assume that the reason to believe operator also satisfies the axioms of positive and negative introspection:

PI $r_i p \rightarrow r_i r_i p$
NI $\neg r_i p \rightarrow r_i \neg r_i p$

Though PI and especially NI are regarded as debatable in the case of knowledge, they seem to be less problematic in the case of the reason to believe operator. First, as noted above, these axioms do not imply that the players actually believe or know that they have or do not have reason to believe something. Second, contrary to knowledge, reason to believe does not imply truth and thus does not enter in tension with NI (Stalnaker 2006: 179). Taken together, the N, K, D, PI and NI axioms yield the so-called KD45 modal logic. The corresponding semantic is a frame $\mathcal{F} = < \Omega, R_1, \ldots, R_n >$ where the accessibility relation R_i satisfies properties of seriality, transitivity and Euclideannes (Bonanno and Nehring 1998; Stalnaker 2006).[15] The semantic counterpart of the reason to believe operator r_i is denoted $\mathrm{R_i}$: $2^\Omega \rightarrow 2^\Omega$ and is defined as follows: $\forall \mathrm{E} \subseteq \Omega$, $\mathrm{R_i E} = \{\omega \mid R_i(\omega) \subseteq \mathrm{E}\}$ where $\mathrm{R_i E}$ is the event that i has reason to believe E. It satisfies the following axioms:

For any events E and F,

(R1) $\Omega = R_i \Omega$
(R2) $R_i(E \cap F) = R_i E \cap R_i F$
(R3) $R_i E \subseteq \neg R_i \neg E$
(R4) $R_i E \subseteq R_i R_i E$
(R5) $\neg R_i E \subseteq R_i \neg R_i E$

[14] This may seem far from being obvious since there is nothing contradictory in having conflicting reasons for believing something or not. For instance, I may consistently have *a* reason to believe that it will rain tomorrow based on the TV's weather forecast but *a* reason to believe that it will not rain based on my experience. Therefore, axiom D entails that reason to believe should be understood as an 'all things considered' statement: if all things considered my reason for believing that it will rain is stronger than my reason for believing it will not rain, the contrary cannot also be true.

[15] The formal statements of these properties are the following: seriality, $\forall \omega \in \Omega$, $\exists \omega'$: $\omega' \in R_i(\omega)$; transitivity, if $\omega' \in R_i(\omega)$ and $\omega'' \in R_i(\omega')$, then $\omega'' \in R_i(\omega)$; Euclideanness, if $\omega' \in R_i(\omega)$ and $\omega'' \in R_i(\omega)$, then $\omega'' \in R_i(\omega')$.

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

Axiom R1 is the semantic counterpart to axiom N and to the fact that all players necessarily have reason to believe the theorems in $F_\theta(G)$. Axiom (R2) is the semantic counterpart to axiom K. It has the interesting implication that for any events E and F such that $E \subseteq F$, if R_iE then R_iF.[16] In other words, each player has reason to believe any logical implication of what he has reason to believe. Axiom R3 corresponds to axiom D and guarantees that each player's reasons to believe are consistent. It does not state however that what one has reason to believe is necessarily true. Axioms R4 and R5 obviously translate axioms PI and NI.

This provides the required basis to define the notion of common knowledge or, more exactly, of *common reason to believe* which plays an essential part in Lewis's theory of conventions. Once again, we can proceed both at the syntactic and the semantic level. For the sake of simplicity, I will only define common reason to believe semantically, though the syntactic counterpart is easy to obtain. Informally, an event E is common knowledge if each person knows E, each person knows that each person knows E, each person knows this, and so on. Similarly, the members of a population *N* have a common reason to believe an event E if each member of *N* has a reason to believe E, each has a reason to believe that everyone in *N* has reason to be believe E, and so on. Formally, we can define semantically the notion of common reason to believe on the basis of the common belief operator R^*. First, we write $R_NE = \cap_{i \in N} R_iE$ for the event that there is mutual reason to believe E in *N* (i.e. everyone in *N* has reason to believe E). Then, common reason to believe E in *N* corresponds to the infinite intersection:[17]

$$R^*E = R_NE \cap R_NR_NE \cap R_NR_NR_NE \cap \ldots$$

Common reason to believe can be also characterized in terms of an accessibility relation R_* where, for every $E \subseteq \Omega$, $R^*E = \{\omega \mid R_*(\omega) \subseteq E\}$ with R_* the transitive closure of $\cup_{i \in N} R_i$. In words, the members of *N* have common reason to believe event E if and only if every $\omega' \in E$ is *reachable* from ω through the *union* of the members' accessibility relations. Note that if there is common reason to believe E there is not common reason to believe $\neg E$ (axiom R3), but also that there is common reason to believe that there is common reason to believe E (axiom R4). However, generally, the fact that there is not common reason to believe E does not entail that there is common reason to believe so (Bonanno and Nehring 1998). Finally, because from the syntactic point of view every event R^*E (resp. $\neg R^*E$) is

[16] The proof is straightforward: since $\omega \in R_iE$ implies that $R_i(\omega) \subseteq E$, it follows from $E \subseteq F$ that $R_i(\omega) \subseteq F$. Hence, we also have $\omega \in R_iF$, which in turn implies $R_iE \subseteq R_iF$.

[17] In a syntactic framework, Sillari (2005: 390) alternatively characterizes common reason to believe as a fixed point: $r^*p \leftrightarrow r_N(p \wedge r^*p)$ with r^* and r_N the common reason to believe and the mutual reason to believe syntactic operators respectively.

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

a theorem of $F_\Theta(G)$, axiom N implies $R_i R^* E$ (resp. $R_i \neg R^* E$) for all persons i.[18]

We now turn to the reasoning part of rule-following behaviour which we capture with the Lewisian concept of *indication*. Lewis (2002 [1969]: 52–53) asked how a given state of affairs may generate among the members of a population a set of higher-order expectations regarding what will unfold:

> Take a simple case of coordination by agreement. Suppose the following state of affairs – call it *A* – holds: you and I have met, we have been talking together, you must leave before our business is done; so you say you will return to the same place tomorrow. Imagine the case. Clearly I will expect you will return. You will expect me to expect you to return. I will expect you to expect me to expect you to return. (…)
>
> What is it about *A* that explains the generation of these higher-order expectations? I suggest the reason is that *A* meets these three conditions:
>
> (1) You and I have reason to believe that *A* holds.
> (2) *A* indicates to both of us that you and I have reason to believe that *A* holds.
> (3) *A* indicates to both of us that you will return.

Then Lewis went on to show that if these three conditions are satisfied, as well as 'suitable ancillary premises regarding our rationality, inductive standards, and background information' (Lewis 2002 [1969]: 53), we can derive an infinite iterative chain of expectations of the kind 'I have reason to believe that you have reason to believe that I have reason to believe … that you will return'. The notion of indication obviously plays a key role in this demonstration. It is not difficult to put Lewis's reasoning in the epistemic and set-theoretic framework developed above. Following Sillari (2005) and Cubitt and Sugden (2003), several axioms characterizing the indication relation need first to be added to the syntax $\mathcal{L}$. In the following, the indication relation is denoted $\Rightarrow_i$ and is intended to mean '_ indicates _ to i'. More specifically, the indication relation captures the relation between two epistemic states through some unspecified kind of practical reasoning. Thus, $p \Rightarrow_i q$ should be read as 'if i has reason to believe p, then *according to some practical reasoning*, he also has reason to believe q'. The point is that one's practical reasoning should include but *is not restricted to* deductive reasoning. On this basis, the following three minimal axioms seem to capture Lewis's concept adequately (to make it clear that the indication relation states a relationship between epistemic

[18] This is formally identical to Cubitt and Sugden's (2014) characterization of an interactive reasoning system as satisfying the conditions of 'awareness', 'authority' and 'attribution'.

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

states, I use the reason to believe operator in the statement of the axioms though this is not formally required):

(I1) $(r_i p \wedge (r_i p \Rightarrow_i r_i q)) \rightarrow r_i q$
(I2) $(p \rightarrow q) \rightarrow (r_i p \Rightarrow_i r_i q)$
(I3) $\forall i, j \in N : ((r_i p \Rightarrow_i r_i (r_j q)) \wedge r_i (r_j q \Rightarrow_j r_j t) \rightarrow r_i p \Rightarrow_i r_i (r_j t)$

The interpretation of these axioms is the following. Axiom I1 is actually the formal translation of the formula 'if i has reason to believe p, then *according to some practical reasoning*, he also has reason to believe q' that I have used to define the indication relation. Axiom I2 reflects the fact that the indication relation does not contradict the relation of logical implication, i.e. deductive reasoning is part of one's practical reasoning. Note that axiom I2 follows from the fact that the indication relation is actually constrained by axiom **N**. Indeed, if $(p \rightarrow q) \in \mathrm{F}_\Theta(G)$, then axiom **N** implies $r_i(p \rightarrow q)$. Axiom I2 then states that the indication relation $\Rightarrow_i$ subsumes the material implication $\rightarrow$. Axiom I3 says that if p indicates to i that j has reason to believe q and if i has reason to believe that q indicates t to j, then p indicates to i that j has reason to believe t.

The R1–R5[19] and I1–I3 axioms permit one to formalize Lewis's conditions for the generation of a common reason to believe in a population. Consider a population N with n persons and denote by $\Rightarrow_N$ the indication relation such 'for each person in N, _ indicates _'. If

L1 $r_N p$
L2 $r_N p \Rightarrow_N r_N (r_N p)$
L3 $r_N p \Rightarrow_N r_N q$

then $r^* q$.

The proof involves an additional axiom which captures Lewis's remark about 'suitable ancillary premises regarding our rationality, inductive standards, and background information'. It states that the members of the population are *symmetric reasoners* (Vanderschraaf 1998; Cubitt and Sugden 2003; Gintis 2009). Basically, if p indicates q to person i and person i has reason to believe that person j has reason to believe p, i and j are symmetric reasoners with respect to p if p indicates to i that j has reason to believe q. That is, i attributes to j the same practical reasoning standards that he endorses for himself, which we can capture by

(SR) $(r_i p \Rightarrow_i r_i q) \rightarrow r_i (r_j p \Rightarrow_j r_j q)$

[19] Or, equivalently, the axioms N, K, D, PI and NI.

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

Combining L2 and L3 with axioms I3 and SR, we obtain[20]

L4 $r_N p \Rightarrow_N r_N(r_N q)$

Combining L2, L4, I3 and SR, we obtain

L5 $r_N p \Rightarrow_N r_N(r_N(r_N q))$

And so on. Moreover, combining L1 and L3 on the basis of axiom I1 gives

L3′ $r_N q$

Similarly, combining L1 and L4 on the basis of axiom I1 leads to

L4′ $r_N(r_N q)$

Since we can proceed in this way for an infinite number of iterations, we have $r^* q$. Semantically, given a model M, the propositions $r_N p$ and $r^* q$ correspond to the events $\mathrm{R_N E}$ and $\mathrm{R^* F}$ such that $M, \omega \models r_N p$ for all $\omega \in \mathrm{R_N E}$ and $M, \omega \models r^* q$ for all $\omega \in \mathrm{R^* F}$ respectively. Following Lewis (2002 [1969]: 56), I will say that E is a basis for the common reason of believing F among the members of the population or, following Cubitt and Sugden (2003), that E is a *common reflexive indicator* of F in the population.

5. INSTITUTIONS AS RULE-GOVERNED GAMES

Consider a game situation described by an epistemic game $\mathcal{G}$. What does it take for the players in $\mathcal{G}$ to follow a rule $\boldsymbol{R}$? Several conditions are required from a Lewisian and Wittgensteinian perspective and lead to the following definition:

Rule-following in $\mathcal{G}$ – An epistemic game $\mathcal{G} : < N, S, \phi, \{u_i\}_{i \in N}, \mathcal{J} >$ is *rule-governed* by a rule $\boldsymbol{R}(\mathcal{G})$ whenever the players implement a strategy profile $\boldsymbol{s}^{\boldsymbol{R}} = (s_1, \ldots, s_n)$ such that three conditions are satisfied:

C1 – Common Understanding in $\mathcal{G}$
C2 – Minimal Awareness
C3 – Minimal Practical Rationality

This section presents and discusses these three conditions. I shall argue that together these three conditions highlight two key insights of Wittgenstein's account of rule-following, namely that rule-following

[20] Combining L3 with axiom SR entails $r_N(r_N p \Rightarrow_N r_N q)$. Combining this result with L2 and axiom I3 entails $r_N p \Rightarrow_N r_N(r_N q)$.

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

depends on the fact that people have a common understanding of the situation they are involved in and that rule-following is fundamentally a community-based practice.

C1 – Common Understanding in $\mathcal{G}$

The first condition is reminiscent of Lewis's definition of convention according to which there must be a common reason to believe that everyone conforms to the convention and that everyone expects everyone to conform to the convention (Lewis 2002 [1969]: 78).[21] It can be stated as follows:

> Proposition C1: Denote E the event that some rule is followed and F the event that the strategy profile $\boldsymbol{s}^{\boldsymbol{R}} = (s_1, \ldots, s_n)$ is implemented in the epistemic game $\mathcal{G}$. The game $\mathcal{G}$ is rule-governed by a rule $\boldsymbol{R}(\mathcal{G})$ only if E is a reflexive common indicator of F in the population.

The axioms stated in the previous section regarding the common reason to believe operator and the indication relation immediately entail several corollaries:

Corollary 1.1: $R_N E$
Corollary 1.2: $R^* F$
Corollary 1.3: Everyone in $\mathcal{G}$ is a symmetric reasoner with respect to E.

Proposition C1 states that there must be a mutual reason to believe that the rule $\boldsymbol{R}$ holds in $\mathcal{G}$ and that this epistemic state must lead to a common reason to believe that some behavioural regularity will be implemented. In some way, the rule *collectively* indicates a behavioural pattern to the members of the population. Indeed, if C1 is the case, then corollaries 1.1 and 1.2 are implied by Lewis's account of common reason to believe. Moreover, as I have shown in the preceding section, symmetric reasoning is a requirement for a mutual reason to believe that an event holds leading to a common reason to believe that another event holds.

Several remarks need to be made here. First, the common reason to believe requirement may seem quite strong as a condition for rule-following behaviour. Indeed, its relevance largely depends (though not

[21] In his definition, Lewis speaks of 'common knowledge' instead of common reason to believe. However, this is improper given that his whole account is not stated in terms of knowledge. Lewis also mentions other conditions constitutive of the convention concept such as the fact that a regularity of behaviour counts as a convention only if there is at least another regularity of behaviour that everyone would prefer to follow provided that there is common reason to believe that everyone follows it. Since I am interested in formalizing *institutions* and since conventions are only a subset of institutions, these conditions are not required in my definition.

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

logically) on the existence of *public events* in the social world, i.e. events which are obviously accessible to everyone in the population when they hold.[22] For instance, Binmore (2008) criticizes Lewis's theory of conventions on the basis of a claim that public events in the social world are probably rare if not non-existent. Moreover, since common reason to believe implies an infinite iterative chain of epistemic states of the kind 'each of us has reason to believe that each of us has reason to believe …', it could be argued that this epistemic state is out of the reach of normal human persons. Regarding the former point, I tend to side with Chwe (2003) who argues that public events are empirically of the utmost importance for the organization of all kinds of human societies. Similarly, Tomasello (2014) has recently argued that the ability to form recursive chains of knowledge (eventually leading to common knowledge) is what distinguishes humans from other animals. The common reason to believe requirement is also a way to make sure that people *actually* follow a rule, i.e. that they behave as they do because of the rule and not – for instance – because of pure randomness. The point is that we are only requiring common *reason* to believe and not actual common belief or knowledge. The only requirement is that the persons could *potentially* support their behaviour and their reasoning by such an iterative chain of epistemic states, in particular if they were asked to justify their behaviour.[23]

A second issue is related to the symmetric reasoning axiom. This is indeed a very strong non-logical assumption about the reasoning abilities of the players. Given the fact that many forms of practical reasoning are available in particular with a variety of inductive standards, the probability that everyone reasons the same way about events seems very small – unless some mechanism is responsible for this 'meeting of minds'. This point is connected with the phenomenon of salience which is discussed by Schelling (1960) and Lewis (2002 [1969]). The latter focused on a particular form of salience, salience because of precedent. Salience is basically the result of the fact that individuals are symmetric reasoners with respect to some event (Hédoin 2014). In the case of rule-following, the rule indicates something to be done to everyone, and there is a common reason to believe so *because* everyone reasons the same way on the basis of some mutually accessible event. Similarly, the fact that some behavioural regularity has occurred in the past provides the members of a population a common reason to believe regarding what should be done in the future only because they basically 'agree' over some kind of

[22] Formally, a public event is defined as an event E such that $R^*E = \{\omega \mid R_*(\omega) \subseteq E\}$ for all $\omega \in E$. In other words, it is an event for which there is *necessarily* a common reason to believe that it holds if indeed it holds.

[23] However, the Minimal Awareness condition requires that people entertain a set of minimal actual beliefs regarding what others are doing. See below.

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

inductive principle. Moreover, there is also a common reason to believe in the fact that everyone is a symmetric reasoner in an informal sense (i.e. not expressed through the common reason to believe operators r_i and R_i).[24] The reason for this lies in the fact that the axiom of symmetric reasoning SR and the common reason to believe operators r^* and R^* are part of the broad theory Θ of G. As a result, all the agents 'know' them because they are used to derive all the formulae in $F_\Theta(G)$.[25]

Therefore, the set-theoretic framework in which we have modelled common reason to believe implies some substantive assumptions regarding how agents reason and about what they know about each other agent's reasoning modes. These assumptions are summarized in Hédoin (2014) by the concept of *common understanding of the (game) situation*. Informally, common understanding of a situation among the members of a population obtains when all persons reason the same way from a given event or state of affairs and this is common knowledge (or common reason to believe). Because common understanding necessarily holds in an epistemic game as formalized above, it is also a constitutive part of any institution. To follow a rule, and thus to behave on the basis of an institution, then presupposes that such a common understanding holds. Interestingly, we find in Wittgenstein's writings about rule-following behaviour a similar idea expressed under the concept of *lebensform* ('form of life') which more or less corresponds to an agreement in judgements between some persons (Wittgenstein 2010 [1953]: §242). As Sillari (2013: 884–885) rightly notes, such an agreement has a natural counterpart in Lewis's theory of common knowledge under the assumption that members of a population share the same inductive standards and that this fact is commonly known among them. For Wittgenstein as well as for Lewis, to follow a rule implies such an agreement regarding the relevant inductive standards.

C2 – Minimal Awareness

Up to now, all the discussion regarding the players' epistemic states has been framed in terms of (common) reason to believe. However, as I have emphasized, reason to believe does not imply actual belief: one may have reason to believe something while not actually believing it. This may be due to cognitive limitations, framing effects or social scripts which may 'activate' or not specific expectations.[26] However, it seems clear that rule-following needs more than mere reason to believe that a rule is followed

[24] See Sillari (2008: 31): 'in the definition of common knowledge itself, there seems to be a problematic requirement of some sort of pre-existing common knowledge'.

[25] Aumann (1987, 1999) makes a similar statement regarding information partitions and the players' prior beliefs within his framework based on S5 modal logic.

[26] See Bicchieri (2005) for a discussion of scripts in the specific case of social norms.

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

in some population. The durability and resiliency of institutions is an indication that at least in some measure, individuals actually expect others to behave in some specific way. The condition of minimal awareness captures this requirement:

> Proposition C2: Denote E the event that some rule is followed and F the event that the strategy profile $\boldsymbol{s}^{\boldsymbol{R}} = (s_1, \ldots, s_n)$ is implemented in the epistemic game $\mathcal{G}$. The game $\mathcal{G}$ is rule-governed by a rule $\boldsymbol{R}(\mathcal{G})$ only if all persons *i* are *minimally aware*, i.e. each *i* actually believes E and F.

The notion of awareness can be formalized on the basis of awareness structures. Following Sillari (2005: 393–397), I will say that an agent *i* actually believes a proposition *p* or an event E if and only if (a) he has a reason to believe *p* or E and (b) he is aware of the proposition or event. To state this definition formally, we need to introduce *n* new epistemic operators a_i such that $a_i p$ means '*i* is aware of *p*'. The semantic counterpart is the operator A_i where

$$A_i E = \{\omega | A_i(\omega) \subseteq E\},$$

with $A_i(\omega)$ an awareness set whose members are all the formulae the agent is aware of at ω. For any person *i*, if an event E is such that both $R_i E$ and $A_i E$, then *i* actually believes E. This is captured by *n* epistemic operators b_i whose semantic counterparts are

$$B_i E = \{\omega | R_i(\omega) \cap A_i(\omega) \subseteq E\}.$$

This is obtained by augmenting the broad theory $F_\Theta(G)$ with a supplementary axiom:

(B) $b_i p \leftrightarrow r_i p \wedge a_i p$

Axiom B is the formal definition of actual belief, i.e. the conjunction of having a reason to believe a proposition and being aware of this proposition. Therefore, condition C2 indicates that a game is rule-governed at a state ω only if both $B_i E \cap B_i F$ for all players *i*. If the belief operator B_i satisfies axiom R2, then this implies that rule-following is constituted by the mutual belief $B_N(E \cap F)$ in the population.[27]

C3 – Minimal Practical Rationality

The last condition we impose on rule-following is standard in a game-theoretic framework since it is shared by the Standard Account of institutions described in section 1.

[27] B_N stands for 'all persons in *N* believe that'.

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

Proposition C3: Denote E the event that some rule is followed and F the event that the strategy profile $\boldsymbol{s}^{R} = (s_1, \ldots, s_n)$ is implemented in the epistemic game $\mathcal{G}$. Assume that all persons i are minimally aware. This implies that each i actually believes that the strategy profile $\boldsymbol{s}^{R}_{-i} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$ is implemented. The game $\mathcal{G}$ is rule-governed by a rule $\boldsymbol{R}(\mathcal{G})$ only if each player i maximizes his expected utility given this mutual belief and his preferences over outcomes represented by the utility function u_i. Formally,

$$D_i s_i \leftrightarrow s_i \in \max s_i [u_i (s_i; \boldsymbol{s}^{R}_{-i}) \mid R_i(\omega) \cap A_i(\omega)](\omega) \text{for all } i \in N$$

$$\text{and for } \omega \in E, \text{and } D_i s_i \text{ the practical statement that } i \text{ does } s_i.$$

Proposition C3 should be uncontroversial. It merely states that each person chooses the strategy that leads to his preferred outcome given his actual belief regarding what others are doing. Note that I have not required the players' beliefs to be accurate. However, since E is a reflexive common indicator of F in the population (and thus everyone is a symmetric reasoner with respect to E) and since everyone is minimally aware (and thus everyone believes that E holds), the modeller knows that the persons' beliefs are correct and indeed this is confirmed by the practice in the population.

The fact that the players in $\mathcal{G}$ are minimally rational implies that rule-following leads to an equilibrium: given the strategy profile $\boldsymbol{s}^{R}$ that is played and everyone's actual (and correct) belief about this, no one has an interest to change his behaviour. This is in accordance with the standard game-theoretic account of institutions which *defines* institutions as equilibria. The relative stability of institutions is clearly related to this feature: because the behavioural pattern corresponds to an equilibrium, it is self-reinforcing. There is an important difference with the Standard Account though: the institution is not *reduced* to this self-reinforcing behavioural pattern. The institution corresponds to the whole fact that the practice is constituted by a rule on the basis of a complex chain of nested and interactive epistemic states between symmetric reasoning agents. Examples range from day to day socially benign practices to economically complex and significant ones. Property rights or complex mechanisms of monetary market exchanges can be described in terms of rule-following behaviour. Consider the former example: the existence of property rights realizes at the behavioural level through a pattern where some persons (the 'owners') actually benefit from the use of different kinds of items, while other people do not interfere with this use. However, the institution of property implies more than that: the behavioural pattern follows from the fact that the members of the population are recognizing some state of affairs (i.e. that there are property rights) and that on this basis, they have a common reason to believe that each person will behave in some way. The

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

recognition of the state of affairs and the ability to infer from this epistemic state a practical conclusion is constitutive of the whole institution of property. The same holds for more complex institutions such as for instance the organization of trade in the community of Maghribi traders during the 11th century (Greif 2006):[28] this organization materialized through a distinctive behavioural pattern which was the product of a complex interaction of beliefs and inferences actually grounded on the fact that the traders were members of the same community.[29]

A Wittgensteinian rationale for both condition C2 and condition C3 can be given. It is partially related with the issue of the origin of the agreement over forms of life (*lebensform*) corresponding to condition C1: where does such an agreement come from? What is the basis for the fact that persons have a common understanding of the situation? Neither Lewis nor Wittgenstein furnishes an answer to this question. Indeed, Wittgenstein seemed to entertain the idea that no such explanation is available (Wittgenstein 1953: §482). However, at the same time, Wittgenstein clearly states that rule-following is fundamentally a *community-based* practice. That is, following a rule is not about interpreting a rule or *thinking* about following a rule (Wittgenstein 1953: §201, §202). Therefore, a person cannot follow a rule 'privately' (Wittgenstein 1953: §202). Following a rule then consists in conforming to the collective practice of some group. But to conform to a collective practice is not only to behave in accordance with some behavioural pattern; it is also to reason along the same standards that the other members. The fact that the community provides the ultimate justification for the use of a particular inductive standard rather than another one is sometimes seen as the 'collectivist solution' to Kripke's sceptical paradox (Bloor 1997). Consider Wittgenstein's famous example of the signpost:

> A rule stands there like a signpost – Does the signpost leave no doubt about the way I have to go? Does it show which direction I am to take when I passed it, whether along the road or the footpath or cross-country? But where does it say which way I am to follow it; whether in the direction of its finger or (for example) in the opposite one? (Wittgenstein 2010 [1953]: §85)

[28] See the appendix for a different but more detailed example based on Greif's work.

[29] Note that instead of persons, players $i = 1, \ldots, n$ in a game can be also interpreted as *social positions* filled by persons. The advantage of this interpretation is that it can account for the fact that within the same institution, a same person may behave differently depending of his position. Within the institution of property rights, owners enjoy free use of their properties and non-owners do not interfere. However, a same and one person will sometimes be in the position of the owner and sometimes in the position of the non-owner and so will change his behaviour accordingly. This implies the recognition of some asymmetry in the game. In the present framework, this recognition is part of the fact that individuals are symmetric reasoners with respect to some event.

This example as well as others is sometimes interpreted as figuring the issue of the *interpretation of the rule*. But this is misguided: the 'right' interpretation is part of the rule – not by a logical necessity but because what the signpost means is determined by an established custom or usage (Wittgenstein 2010 [1953]: §198) belonging to a group or a community. Following the rule (e.g. behaving in a particular way with respect to a signpost) *consists in* the fact that everyone reason the same way on the basis of a given state of affairs. No interpretation beyond the rule is required. A similar statement is made by Lewis (2002 [1969]: 61): 'So if a convention, in particular, holds as an item of common knowledge, then to belong to the population in which that convention holds – to be party to it – is to know, in some sense, that it holds.' A rule necessarily belongs to a collective or a community and to be a member of that community is to believe (and thus to be aware) that the rule holds. In other words, what makes two persons belong to the same community is the fact that they have reason to believe that the same rule(s) hold, and that they share a form of life, i.e. a set inductive standards. This leads to an actual belief regarding what other members of the community will be doing in some given situation. Ultimately, this is through this relationship between this actual belief and the individuals' minimal practical rationality that rule-following and institutions are essentially community-based.

6. CONCLUSION

The main point of this paper has been to propose a theory of rule-following in a game-theoretic framework. The result is a formalization of institutions as epistemic games that significantly departs from the standard game-theoretic account of institutions, where the latter are defined as mere behavioural regularities.

I would like to finish this article by briefly considering one objection that could be formulated against my account. If the rejection of the standard account of institutions discussed in the first section is understood as a conceptual and philosophical critique, then one may concur with Ross (2014) that such a critique is irrelevant from the scientific point of view. Indeed, this is a point with which I largely agree: scientists create and use (as they should) concepts in an essentially instrumental perspective to enhance our understanding of the world as well as our ability to interact with it. A conceptual critique whose basis is the fact that scientists use a concept in the sense that does not correspond to some 'folk ontology' is mostly irrelevant and will generally be rightly disregarded by scientists. However, the argument developed in this paper should not be understood as a conceptual critique. The point is not that behavioural regularities should not be called 'institutions' or that the 'essence' of institutions is something different. This sort of metaphysical consideration

has little value from a scientific, if not philosophical point of view. Rather, what has been emphasized is that the game theorists' concept of institutions does not capture a mechanism that plays a significant role in the social world. This social mechanism is what lies behind rule-following behaviour. My goal is thus not to redefine the concept of institutions but to make the case for a finer theoretical partition of the mechanisms ruling the social world.

ACKNOWLEDGEMENTS

A previous version of this paper has been presented at the Economic Philosophy Seminar of the GREQAM research centre (Aix-Marseille University, France) in June 2014 and at the 2nd Economic Philosophy Conference (Strasbourg, France) in October 2014. I would like to thank the two anonymous referees of the journal for their useful and stimulating comments which have contributed to enhance the quality of the paper. I also thank Richard Bradley for his editorial advice.

REFERENCES

Aoki, M. 2001. *Toward a Comparative Institutional Analysis*. Cambridge, MA: MIT Press.

Aumann, R. J. 1976. Agreeing to disagree. *Annals of Statistics* 4: 1236–1239.

Aumann, R. J. 1987. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* 55: 1–18.

Aumann, R. J. 1999. Interactive epistemology I: knowledge. *International Journal of Game Theory* 28: 263–300.

Aumann, R. J. and J. H. Dreze. 2008. Rational expectations in games. *American Economic Review* 98: 72–86.

Bacharach, M. 1987. A theory of rational decision in games. *Erkenntnis* 27: 17–55.

Bacharach, M. 1994. The epistemic structure of a theory of a game. *Theory and Decision* 37: 7–48.

Bicchieri, C. 2005. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

Binmore, K. 2008. Do conventions need to be common knowledge? *Topoi* 27: 17–27.

Bloor, D. 1997. *Wittgenstein, Rules and Institutions*. New York, NY: Routledge.

Bonanno, G. 2002. Modal logic and game theory: two alternative approaches. *Risk, Decision and Policy* 7: 309–324.

Bonanno, G. and K. Nehring. 1998. Assessing the truth axiom under incomplete information. *Mathematical Social Sciences* 36: 3–29.

Chwe, M. S-Y. 2003. *Rational Ritual: Culture, Coordination, and Common Knowledge*. Princeton, NJ: Princeton University Press.

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

Cubitt, R. P. and R. Sugden. 2014. Common reasoning in games: a Lewisian analysis of common knowledge of rationality. *Economics and Philosophy* 30: 285–329.

Fudenberg, D. and D. K. Levine. 1998. *The Theory of Learning in Games*. Cambridge, MA: MIT Press.

Gintis, H. 2009. *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton, NJ: Princeton University Press.

Greif, A. 2006. *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade*. Cambridge: Cambridge University Press.
Greif, A., P. Milgrom and B. R. Weingast. 1994. Coordination, commitment, and enforcement: the case of the merchant guild. *Journal of Political Economy* 102: 745–776.
Hédoin, C. 2014. A framework for community-based salience: common knowledge, common understanding and community membership. *Economics and Philosophy* 30: 365–395.
Hédoin, C. 2015. Accounting for constitutive rules in game theory. *Journal of Economic Methodology*, forthcoming.
Hindriks, F. 2009. Constitutive rules, language, and ontology. *Erkenntnis* 71: 253–275.
Hindriks, F. and F. Guala. 2015. Institutions, rules, and equilibria: a unified theory. *Journal of Institutional Economics* 11: 459–480.
Lahno, B. 2007. Rational choice and rule-following behavior. *Rationality and Society* 19: 425–450.
Lewis, D. K. 2002 [1969]. *Convention: A Philosophical Study*. Oxford: Blackwell.
Maynard Smith, J. 1982. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.
Milgrom, P. R., D. C. North and B. R. Weingast. 1990. The role of institutions in the revival of trade: the law merchant, private judges, and the Champagne fairs. *Economics and Politics* 2: 1–23.
Perea, A. 2012. *Epistemic Game Theory: Reasoning and Choice*. New York, NY: Cambridge University Press.
Rawls, J. 1955. Two concepts of rules. *Philosophical Review* 64: 3.
Ross, D. 2009. Reply to Hands: on the Robbins-Samuelson argument pattern. *Journal of the History of Economic Thought* 31: 93–103.
Ross, D. 2014. *Philosophy of Economics*. London: Palgrave Macmillan.
Schelling, T. C. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.
Schotter, A. 2008 [1981]. *The Economic Theory of Social Institutions*. Cambridge: Cambridge University Press.
Searle, J. R. 1995. *The Construction of Social Reality*. New York, NY: Simon and Schuster.
Searle, J. R. 2010. *Making the Social World: The Structure of Human Civilization*. New York, NY: Oxford University Press.
Sillari, G. 2005. A logical framework for convention. *Synthese* 147: 379–400.
Sillari, G. 2008. Common knowledge and convention. *Topoi* 27: 29–39.
Sillari, G. 2013. Rule-following as coordination: a game-theoretic approach. *Synthese* 190: 871–890.
Skyrms, B. 1996. *Evolution of the Social Contract*. Cambridge: Cambridge University Press.
Smit, J. P., F. Buekens and S. du Plessis. 2011. What is money? An alternative to Searle's institutional facts. *Economics and Philosophy* 27: 1–22.
Smit, J. P., F. Buekens and S. du Plessis. 2014. Developing the incentivized action view of institutional reality. *Synthese* 191: 1813–1830.
Stalnaker, R. 2006. On logics of knowledge and belief. *Philosophical Studies* 128: 169–199.
Sugden, R. 2005. *The Economics of Rights, Cooperation and Welfare*. London: Palgrave Macmillan.
Tomasello, M. 2014. *A Natural History of Human Thinking*. Cambridge, MA: Harvard University Press.
Vanberg, V. J. 2004. The rationality postulate in economics: its ambiguity, its deficiency and its evolutionary alternative. *Journal of Economic Methodology* 11: 1–29.
Vanderschraaf, P. 1998. Knowledge, equilibrium and convention. *Erkenntnis* 49: 337–369.
Wittgenstein, L. 2010 [1953]. *Philosophical Investigations*. Malden, MA: Wiley-Blackwell.

APPENDIX 1. AXIOMS, POSTULATES AND PROPOSITIONS FOR $\mathcal{J}$

This appendix summarizes the axioms, postulates and propositions constitutive of the broad theory $\mathcal{J}$ of a game G whose epistemic counterpart $\mathcal{G}$ is rule-governed by a rule $\boldsymbol{R}(\mathcal{G})$.

For any game $G = < N, S, \phi, \{u_i\}_{i \in N} >$, the broad theory $\mathcal{J}$ of G provides a full description of how the game is played and of the players' knowledge, beliefs and reasoning modes. The broad theory $\mathcal{J}$ is constituted of a theory of rational play Θ and a set NL of non-logical axioms and postulates for Θ. The resulting tuple $< N, S, \phi, \{u_i\}_{i \in N}, \mathcal{J} >$ corresponds to an epistemic game $\mathcal{G}$.

The formalization of an institution as a rule-governed game in the paper depends on a broad theory $\mathcal{J}$ with the following axioms, postulates and propositions:

Theory of rational play Θ

The theory of rational play Θ is a formal language defined by

- A set of *practical statements* of the kind $D_i s_i$ ('i does s_i'), for all $i = 1, \ldots, n$.
- The connectives $\neg$ and $\vee$.
- A set of n reason to believe operators r_i ($i = 1, \ldots, n$) satisfying a set of logical axioms:
  - N: $p \rightarrow r_i p$
  - K: $r_i p \wedge r_i(p \rightarrow q) \rightarrow r_i q$
  - D: $r_i p \rightarrow \neg r_i \neg p$
  - PI: $r_i p \rightarrow r_i r_i p$
  - NI: $\neg r_i p \rightarrow r_i \neg r_i p$
- A set of n awareness operators a_i ($i = 1, \ldots, n$).
- A set of n belief operators b_i ($i = 1, \ldots, n$) defined as $b_i p \leftrightarrow r_i p \wedge a_i p$.
- A common reason to believe operator $r^* p \leftrightarrow r_N(p \wedge r^* p)$.

Non-logical axioms and postulates NL

We assume that players have a common reason to believe that they play game G:

(CRBG) $r^* G$

For all $i, j \in N$, the indication relation $\Rightarrow_i$ satisfies

(I1) $(r_i p \wedge (r_i p \Rightarrow_i r_i q)) \rightarrow r_i q$
(I2) $(p \rightarrow q) \rightarrow (r_i p \Rightarrow_i r_i q)$
(I3) $\forall i, j \in N: ((r_i p \Rightarrow_i r_i(r_j q)) \wedge r_i(r_j q \Rightarrow_j r_j t) \rightarrow r_i p \Rightarrow_i r_i(r_j t)$
(SR) $\forall i, j \in N, \exists p: (r_i p \Rightarrow_i r_i q) \rightarrow r_i(r_j p \Rightarrow_j r_j q)$

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

			Borrower	
		Reimburse		Do not Reimburse
	Lend	$1+i;g$		$0;G$
Lender				
	Do not lend	$r;0$		$r;0$

FIGURE 1. The Loan game.

If we write p for 'rule $\boldsymbol{R}$ holds' and q for 'the strategic profile $\boldsymbol{s}^{\boldsymbol{R}} = (s_1, \ldots, s_n)$ is implemented', then:

Condition C1 (Common Understanding) entails

(C1a) $\forall i$: $r_i(r_N p)$
(C1b) $r^* q$
(C1c) $\forall i \in N$: $r_i(r_N p) \Rightarrow_i r_i(r^* q)$

Condition C2 (Minimal Awareness) corresponds to

(C2a) $\forall i \in N$, $b_i p \wedge b_i q$

Condition C3 (Minimal Practical Rationality) corresponds to

(C3a) $\forall i \in N$: $D_i s_i \leftrightarrow s_i \in \max s_i[u_i(s_i; \boldsymbol{s}^{\boldsymbol{R}}{}_{-i}) \mid \mathrm{R_i}(\omega) \cap \mathrm{A_i}(\omega)]\ (\omega)$

APPENDIX 2. AN EXAMPLE: THE LOAN GAME

Consider an interaction between a lender and a borrower concerning the bargaining over a loan contract and that we formalize by the following 'Loan Game'.[30]

Suppose that the following inequalities hold:

$r > 0$ (The lender's payoff if he does not lend are strictly positive)
$1 + i > r$ (The lender prefers to lend if the borrower reimburses)
$G > g > 0$ (The borrower prefers to not reimburse if the lender lends and it is better for the borrower to obtain a loan than to not obtain it)
$G < g + i + l$ (It is socially better that the loan is made and reimbursed)

Assume that both the lender and the borrower actually believe that they are playing the Loan Game. Then, the condition C3 (Minimal Practical Rationality) obviously entails that the borrower will not reimburse the loan which in turn entails that the lender will not lend. Suppose however that the borrower will have to interact in the next period with another lender with a strictly positive and not

[30] This example is based on Avner Greif's (2006) model of the community responsibility system. I intentionally use a very simplified version to make my point transparent.

https://doi.org/10.1017/S0266267116000043 Published online by Cambridge University Press

too low probability p. The borrower and all the potential lenders form a population P. Now an institution formalized by the Loan Game and leading to the conclusion of a loan contract could be the following.

Denote E the event 'rule ***R*** holds' and where ***R*** corresponds to 'if Borrower has not reimbursed a lender at period t, then Borrower never gets a loan for any period $t^* > t$; if Borrower has reimbursed a lender at period t, then Borrower gets a loan for any period $t^* > t$'.[31] Denote F the event 'Lenders lend as long as Borrower reimburses, Borrower reimburses' (F thus consists in a profile of conditional strategies). Suppose that E is a common reflexive indicator of F among the members of P, i.e. provided that there is a mutual reason to believe E in P, then there is a common reason to believe F in P. Assume that condition C2 (Minimal Awareness) holds for F if it holds for E (hence everyone actually believes F if everyone actually believes E). Now, if the probability that the borrower will meet another lender is sufficiently high, the fact that F is mutually believed will induce the borrower to reimburse the loan at t. Indeed, it is clear that condition C3 entails that any lender should lend if F is actually believed. Then, for the borrower, reimbursing is rational given his actual beliefs if $g/(1-p) > G$ with p the probability that the borrower meets a lender at the next period.

The whole institution consists in the networks of reasons to believe, actual beliefs and reasoning standards over events E and F as well as in the behavioural pattern it leads to. The practice described by the game depends on a (constitutive) rule ***R***, i.e. the practice is rule-governed by the rule ***R***. Note that the rule does not reduce to a statement or to an event; rather, it corresponds to an epistemic and practical nexus where some states of beliefs over the fact that the rule holds entails a practical conclusion over what one should do. In this sense, the institution depends on a (set of) constitutive rule(s). Furthermore, it works on the basis of the fact that the players share some inductive standards (condition C1). On a Wittgensteinian reading, this is due to the fact that they are members of the same community.

BIOGRAPHICAL INFORMATION

Cyril Hédoin is Full Professor of Economics at the University of Reims Champagne-Ardenne, France. He has recently published papers in the *Erasmus Journal for Philosophy and Economics*, the *Journal of Economic Methodology* and *Economics and Philosophy*. His academic work essentially belongs to the philosophy of economics and to institutional economics. His recent publications deal with the issue of the nature of rules and institutions and with their formalization within a game-theoretic framework.

[31] It is important to note that the players need not have (and likely do not have) this specific content in mind when following the rule. Indeed, the Wittgensteinian stance adopted in this paper means that I assume externalism about mental states: the semantic content of mental states such as beliefs is not intrinsic to the being having these states. In other words, it is irrelevant (or meaningless) to ask what the players have 'in their minds' when they follow the rule. What counts is that the players know that a rule holds and that on this basis they follow some behavioural pattern. It is only from the latter that we can infer the semantic content of the players' mental states.