Laboratoire REGARDS (EA 6292)
Université de Reims Champagne-Ardenne

Working paper n° 6-2014

Against the Eliminativist View of Institutions in Economics: Rule-Following and Game Theory

Cyril Hédoin*

* Professeur des Universités en sciences économiques, REGARDS (Université de Reims Champagne-Ardenne)

Abstract
This paper argues that most game-theoretic accounts of institutions in economics are eliminativist: they reduce institutions to behavioral patterns the players are incentivized to implement. However, because institutions do not explain the behavior of the agents, they can be eliminated as a concept from the scientific explanation. An alternative account linking institutions to rule-following behavior in a game-theoretic framework is developed. Institutions are defined as epistemic games where the players have a common prior over the state space. I show that the assumptions of common prior and of a commonly known state space in an epistemic game are the same as to assume that the members of a community have a common understanding of a situation. This common understanding has a strong similarity with Ludwig Wittgenstein's concept of *lebensform*.

Mots clés : institutions, rule-following, epistemic game theory, common understanding, Wittgenstein

Laboratoire d'Economie et
Gestion REGARDS (EA 6292)
Université de Reims
Champagne-Ardenne
UFR de sciences économiques,
sociales et de gestion
57B Rue Pierre Taittinger
51096 Reims

Directeur : **Martino Nieddu**

Responsable de l'édition des
working papers **: Cyril Hedoin**

WP 6-2014

Against the Eliminativist View of Institutions in Economics: Rule-Following and Game Theory

Cyril Hédoin[*]

REGARDS (EA 6292) – University of Reims Champagne-Ardenne, France

This version: 30/09/2014

Abstract: This paper argues that most game-theoretic accounts of institutions in economics are eliminativist: they reduce institutions to behavioral patterns the players are incentivized to implement. However, because institutions do not explain the behavior of the agents, they can be eliminated as a concept from the scientific explanation. An alternative account linking institutions to rule-following behavior in a game-theoretic framework is developed. Institutions are defined as epistemic games where the players have a common prior over the state space. I show that the assumptions of common prior and of a commonly known state space in an epistemic game are the same as to assume that the members of a community have a common understanding of a situation. This common understanding has a strong similarity with Ludwig Wittgenstein's concept of *lebensform*.

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

Word count: 13000

[*] Paper prepared for the 2nd international conference "Philosophy of economics", Strasbourg, 9-10/10/2014. A first version has been presented at the "Economie-Droit-Philosophie" seminar of the GREQAM, University of Aix-Marseille, 11/06/2014. I thank the participants for their useful and stimulating remarks.
Email: cyril.hedoin@univ-reims.fr

Against the Eliminativist View of Institutions in Economics: Rule-Following and Game Theory

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 is Andrew Schotter's *The Economic Theory of Social Institutions* (Schotter 1981). Since then, many authors have developed this strand of research further (*e.g.* (Aoki 2001); (Sugden 2005); (Young 1998)). Avner Greif's (Greif 2006) insightful historical and theoretical study of the institutional foundations of economic development may be seen as the culmination of a research program that started some thirty years ago.

In two recent articles, J.P. Smit, Filip Buekens and Stan du Plessis ((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. The incentivized action view is indeed nothing but a general and literal description of how institutions are accounted for in a game-theoretic framework: institutions are viewed as behavioral pattern emerging on the basis of the players' preferences and beliefs. I shall argue in this article that this approach which I have called elsewhere the "standard game-theoretic account of institutions", or Standard Account for short (Hédoin 2014b), underlies an *eliminativist* view of institutions: institutions do not really exist in the sense that they do not play any logical or causal role in explaining the agents' behavior. They are merely behavioral patterns which summarize the history of plays in a strategic interaction.

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 behavioral 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 than 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 1953).

The article is organized as follows. In the second section, I present with some details the standard game-theoretic account of institutions. I characterize the conception of institutions that it underlies as "eliminativist" by analogy with the eliminativist paradigm in the philosophy of mind in the third section. The fourth section briefly considers the notion of rule-following in Ludwig Wittgenstein's and David Lewis' writings. The fifth section develops a rule-following account in game theory where I define institutions as epistemic games. The sixth section considers the role of communities in fostering the common understanding which is constitutive of institutions. A seventh section briefly concludes by examining two potential objections to my account.

2. The Standard Game-Theoretic Account of Institutions

Economists are generally not very explicit regarding what institutions really are.[1] However, particularly since the beginning of the development of the so-called "new institutional economics" in the 1970's, it has been recognized that "institutions matter" both at the microeconomic level of the firms and consumers (why do firms exist? Why is trade organized in such or such a way?) and at the more macroeconomic level of economic development (why do some economies enjoy economic growth but not others?). The focus on institutions in economics is now so well established that many empirical studies try to measure their impact on economic variables such as growth or inequalities. But because they remain conceptually unclear, institutions are yet to be fully integrated at the theoretical level.

Game theory has been for at least the last thirty years the main tool on the basis of which the conceptual and theoretical investigations around the concept of institution have been led. However, the game theorists' sudden interest in institutions did not rise from some empirical concerns but rather as a necessity to go beyond a theoretical bypass. Indeed, the growing emphasis on institutions in a game-theoretic framework largely finds its roots in the well known equilibrium selection problem, which itself gives rise to the indeterminacy problem. If we suppose (as is common among game theorists) that any solution concept for a game must correspond to some equilibrium, *i.e.* an outcome where no player has an incentive to change his strategy, then it appears that many games have more than one equilibrium.[2] In this case, game theory fails to predict what the outcome implemented by the players in the game will be or, in a more normative stance, fails to make a strategic recommendation for each player. As is well known, this problem is even more significant if the players have a common knowledge of their rationality (Sugden 1991). Even more important, the failure of the game theorist to predict (or recommend) the outcome of the game extends to the players themselves: if rational

[1] There are exceptions though, *e.g.* Hodgson (2006).

[2] This problem is of course particularly acute in the case of the Nash equilibrium solution concept. But it is also relevant for some "stricter" solution concepts such as sequential equilibrium or perfect equilibrium.

players know game theory as well as the game theorist does, then they cannot predict what others will do and hence what they should do. They face therefore an indeterminacy problem: the best way to play into the game is indeterminate and game theory is unhelpful as an algorithm to choose optimally.

Institutions can be inserted in a game-theoretic framework in several ways. For instance, Aoki (2001, 4-9) argues that there are at least three views of institutions in game theory. The first view refers to institutions as players in a game: in this case, institutions are essentially organizations with a well-defined preference ordering or objective function which take decisions on the basis of some rationality assumptions. The second view, originally due to Douglass North (1990), famously pictures institutions as the "rules of the game in a society or… the humanly devised constraints that shape human interaction" (North 1990, 3-4). Over the last two decades, this view has essentially been taken over by the so-called mechanism design literature. For instance, following Hurwicz's (1996) game-theoretic account, an institution is equivalent to a *game form, i.e.* a triple defined by a set of players, a set of pure strategies and an outcome function. Therefore, in this sense, institutions are rules which determine who is entitled to play, what each player is technologically and legally entitled to do and what happens in each possible scenario. To count as an institution, Hurwicz adds that these rules must be enforceable, which he means by this that they must lead to a Nash equilibrium. As I will briefly highlight in the sixth section, the mechanism design approach to institutions shares some important properties with the rule-following account I develop in the fifth section. But it falls short from being the most popular way to formalize institutions in a game-theoretic framework.

Indeed, the most fashionable game-theoretic approach to institutions is what I call elsewhere the "Standard Account" and corresponds to Aoki's third view, according to which institutions are equilibria in games. In other words, institutions are identified as self-enforcing patterns of behavior and/or consistent epistemic attitudes.[3] Formally, in the Standard Account, 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 last game-theoretic view of institutions actually encompasses a great variety of modeling 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[4] 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

[3] By "epistemic attitudes", I refer to the fact that game theorists are sometimes interested in identifying equilibria in beliefs, which are also called Bayesian equilibria. Since a Nash equilibrium can also be characterized as a situation where the players hold mutually consistent beliefs, it is frequent to speak of a Nash-Bayes equilibrium. In this paper, the notion of equilibrium is always understood in this latter, encompassing, sense.

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

sophisticated theoretical and historical accounts of informal and formal institutions sustaining and organizing trade and other kinds of economic exchange.[5] Between these two extremes, 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). Schotter (1981) and Young (1998) give two of the most important game-theoretic accounts of institutions where the latter are seen as the aggregation of the choices of mildly rational players learning how to play on the basis of their (possibly imperfect) knowledge of the history of the game. For instance, Young (1998)[6] formalizes the strategic pattern in an infinitely repeated n players game where players are randomly paired over time in a fictitious play dynamic. Informally, it is assumed that each player observes a sample of the history of plays in the game and plays his best reply (*i.e.* the strategy maximizing his expected utility). Crucially, each player makes the assumption that the other players will play according to the strategy frequencies found in the sample. Young then shows mathematically that this leads to the emergence of behavioral patterns corresponding to empirically significant institutions such as fairness and bargaining norms.

Even though the formalism corresponding to the Standard Account is far from being homogenous, a generic formal framework can nevertheless be found to hold in the background, as is shown by Aoki (2001, 186-194). To make this framework explicit will help to understand why this game-theoretic account of institutions is eliminativist. The core of the Standard Account corresponds to what is called a "game form", *i.e.* a triple G: $< N, S, \phi >$. $N = \{1, \ldots, n\}$ is the set of players. $S = S_1 \text{ x } \ldots\text{x } S_n$ 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: $S \rightarrow X$ mapping any strategy profile $(s_1, \ldots, s_n)$ into an outcome $x \in X$ is called a *consequence function* and is the formal expression of the exogenously given rules of the game.[7] From this generic structure, we can define a generic mechanism representing the choice pattern that will unfold in any game G repeated infinitely among a population of n players. Denote this repeated game by $\boldsymbol{G}$. Consider that each player follows a private strategy-choice rule r_i: $X \rightarrow S_i$ such that:

(1) $s_i(t + 1) = r_i(x(t))$

Expression (1) simply states that a player i's choice of a strategy at time t+1 is a function of what happened in the previous iteration. A more general rule is to allow i's choice at t+1 to be a function of the whole history $h(\boldsymbol{G}; t) = (x(t), x(t\text{-}1), \ldots, x_0)$ of $\boldsymbol{G}$ at time t, with x_0 the null

[5] The work of Avner Greif (2006) is probably the most significant example of this approach. See also Greif, Milgrom, and Weingast (1994) and Milgrom, North, and Weingast (1990).

[6] For a presentation of Schotter's game-theoretic account of institutions and how it relates to the Standard Account, see Hédoin (2014b).

[7] The function ϕ is determined by the whole institutional and social environment, in which the game takes place. The key point is that it is an exogenous but also objective feature *from the players' point of view*.

history. Let H_i be the set of all potential histories for a player *i*. Then the more general private strategy-choice rule r_i: $H_i \rightarrow S_i$ is:

(2) $s_i(t + 1) = r_i(h_i(\boldsymbol{G}, t))$

The function r_i may represent a large set of psychological, cognitive or social mechanisms that I will assume for the moment are undefined.[8] Indeed, in the incentivized action view of Smith et al.'s ((2011); (2014)), what determines r_i is considered to be of secondary importance if not totally irrelevant. I shall later argue that in a rule-following perspective, what lies behind r_i is on the contrary essential to understand the nature of institutions. On the basis of (2), we can determine an aggregative rule of transition ***r*** from one state (a component of *X*) to another through time. This rule defines a transition function *F*: $X \rightarrow X$:

(3) $x(t + 1) = \boldsymbol{r}(\phi(x(t))) = F(x(t))$

As noted above, game theorists put a major emphasis on equilibrium reasoning. The notion of equilibrium finds a natural expression in this framework since an equilibrium simply corresponds to a rest point in the dynamic such that the system defined by ***G*** stays infinitely in the same state. In other words, any state x^* such that $x^* = F(x^*)$ is an equilibrium. The actual dynamic (and the stability of the rest points) depends on the way the players make choices in the game. In turn, that depends on the nature and the degree of rationality with which the players are endowed.

Consider first a pure evolutionary framework akin to the one developed by Skyrms (1996) or Sugden (2005). A specific aggregative transition rule then is the *replicator dynamic*. It states that the increase (or the decrease) in the proportion of the followers of a given strategy in the population is a linear function of the difference between the (expected) payoff of this strategy and the mean (expected) payoff of the population. To formally state this transition rule, it is first necessary to add to the game form *G* a component determining how each player subjectively values each potential physical outcome. For each player, we thus define a utility function u_i: $X \rightarrow \Re$ mapping each outcome into some real number. Now, if we write $p(s; t)$ the proportion of followers of strategy *s* in the population at time *t*, a general expression of the replicator dynamics is

(4) $p(s; t + 1) = p(s; t) + f(u(s; t); \boldsymbol{u}(t))$,

where $u(s; t)$ is the expected payoff brought about by strategy *s* at *t* and $\boldsymbol{u}(t)$ is the mean expected payoff in the population at *t*. Accordingly, the function *f* is such that

$f(u(s; t)) > 0$ if $u(s; t) > \boldsymbol{u}(t)$

$f(u(s; t)) = 0$ if $u(s; t) = \boldsymbol{u}(t)$

$f(u(s; t)) < 0$ if $u(s; t) < \boldsymbol{u}(t)$

[8] Indeed, r_i may represent a more or less sophisticated learning rule such as Young's best reply, as well as very basic routine-based behavior.

On the basis of the transition dynamic defined by the function *f*, we can identify an asymptotically stable equilibrium as a rest point $\boldsymbol{p^*}$ in the dynamic such that for any initial position in the neighborhood of $\boldsymbol{p^*}$, $\lim_{t\to\infty} \boldsymbol{p}(t) = \boldsymbol{p^*}$, with $\boldsymbol{p}$ a vector of strategy distribution in the population. In other words, in such an evolutionary framework, an institution is identified to an evolutionary equilibrium, *i.e.* a dynamically stable strategy distribution in the population.

A similar idea holds for approaches based on repeated games and the subgame perfect equilibrium solution concept. Here it is assumed that at time $t = 0$ each player chooses a strategy which consists of an exhaustive contingent plan assigning an action at each information set of the game tree. Such a strategy (the so-called "grim trigger strategy") may be for instance to cooperate until the other player defects, and then to defect indefinitely. Subgame perfection restricts the set of equilibrium strategies to those which depend on credible commitments, *i.e.* strategies in which all actions are optimal at the given information set even for those information sets that are never actually attained. The point here is that, contrary to the evolutionary case, agents are foresighted.[9] Formally, a strategy s is thus a comprehensive plan of future actions contingent on any potential history of the game $h_i(\boldsymbol{G}; t) \in H_i$. A strategy is thus formally identical to the strategy-choice rule r_i except for the fact that the rule determines actions (one move at a particular information set) but not strategies (a comprehensive plan of actions), *i.e.* r_i: $H_i \rightarrow S_i$. It follows that the transition function F: $X \rightarrow X$ is simply determined by the vector of players' strategies $\boldsymbol{s} = \{s_1, \ldots, s_n\}$. For any history $h(\boldsymbol{G}, t)$ and its associated outcome $x(t)$ at t, a *subgame* corresponds to all the moves made from that period onwards. For any period $\tau > t$ and any history $h(\boldsymbol{G}, t)$, the strategic path of the game in the corresponding subgame will be written $\boldsymbol{x}(\tau : h(\boldsymbol{G}, t))$.

In repeated games, two conditions regarding the rationality of the players are most of the time imposed: a) each player i must form a consistent expectation σ_i: $H_i \rightarrow S_{-i}$ regarding the strategy choice of others; b) each player must maximize his expected utility in each and every subgame of the larger game $\boldsymbol{G}$ given his expectation. Formally, these two conditions can be written respectively:

$$(5) \qquad \sigma_i(\tau : h(\boldsymbol{G}, t)) = \boldsymbol{x}_{-\boldsymbol{i}}(\tau : h(\boldsymbol{G}, t)),$$

for any player $i = 1, \ldots, n$, any $t > 0$, and any history $h \in H$.

$$(6) \qquad s_i \in \max_{s_i} \textstyle\sum_{\tau > t} \delta^{\tau - t} u_i(s_i(\tau : h(\boldsymbol{G}, t)), \sigma_i(\tau : h(\boldsymbol{G}, t)),$$

for all player $i = 1, \ldots, n$, all $t > 0$, and all history $h \in H$.

If these two conditions are satisfied, then it is easy to check that every strategic path $\boldsymbol{x}(\tau : h(\boldsymbol{G}, t))$ is subgame perfect for any history h and any period t. As a consequence, the whole strategic path $\boldsymbol{x}(\tau : x_o)$ is necessarily a subgame perfect equilibrium for any $\tau > 0$. As Aoki (2001, 189) puts it, in this situation "it is not beneficial for any agent to unilaterally deviate

[9] In repeated game models, it is moreover generally assumed that agents discount future payoffs at some positive rate. This is captured by the discount factor δ which enters into the utility function of each player.

from the specified strategy at any moment in time, whatever the history up to that point may be, and thus the chosen *action-choice rules become self-enforcing*" (emphasis in original).[10] A subgame perfect equilibrium is thus nothing but an equilibrium of contingent strategies. An interesting consequence is that such an equilibrium will generate a regular pattern of actions or, in other words, *ergodic histories* of play. This recurrent pattern of plays corresponds to what most game theorists using repeated games call "institutions".[11]

3. The Standard Account as an Eliminativist View of Institutions

This section links the standard game-theoretic account of institutions characterized above with Smit et al.'s ((2011); (2014)) "incentivized action view of institutional reality". My point is to make it clear that Smit et al.'s account, though less formal, is conceptually equivalent to the Standard Account. In the process, I will characterize by means of an analogy with the philosophy of mind literature the corresponding view of institutions as an *eliminativist* one. Basically, I shall argue that both in the Standard Account and in 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 behavioral 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* endeavor. As a scientific and explanatory concept referring to some part of the social reality, it can simply be eliminated from the economist's vocabulary.

Smit et al.'s "incentivized action view of institutional reality" builds on a critique of John Searle's theory of institutional fact ((1995); (2010)). Searle argues that the nature of social reality relies on institutional facts where people collectively accept to grant objects or persons with a peculiar status with a set of associated deontic powers. More precisely, according to Searle, all institutional facts are in the form of "this X counts as Y in Z" where the "count as" locution indicates the status and the deontic powers that are collectively recognized to the object or person X (in conditions Z) by some collective. For instance, an important institutional fact extensively studied by Searle is money: a piece of paper counts as money in certain circumstances because people collectively recognize that this piece of paper has the power to buy things. An important aspect of Searle's account of institutional facts is that social reality depends on the ability of individuals to develop a form of collective intentionality, *i.e.* collective intentional states (desires, beliefs, intentions, …) of the form "*we* desire/believe/intend that…". Smit et al. (2011) particularly criticize this last feature of Searle's account: while remaining agnostic regarding the possibility of collective intentionality as defined by Searle, they argue that institutional facts do not depend on collective intentionality. However, the main point of their dispute is what they call "Searle's strong irreducibility thesis", namely the fact that Searle's account needs to postulate the

[10] Aoki speaks of "action-choice rules", which are the same as the strategy-choice rules in the text.

[11] In some cases (*e.g.* Greif (2006)), institutions refer to the pattern of play *and* the associated consistent belief distribution.

existence of irreducible institutional facts, *i.e.* facts that cannot be accounted for in other ways than in the very same institutional terms. Because of this, the authors blame Searle's account for its circularity. For instance, the institutional fact of money cannot be accounted for without making a reference to the institutional fact of money: a piece of paper counts as money because some members of a collective collectively accept that this is the case. This self-reference must be eliminated from any theory of institutions, or so the authors pretend.

My point here is not to evaluate the relevance of Smit et al.'s critique of Searle's theory of institutional facts. My interest rather lies in the alternative account of institutions and institutional facts they propose. The latter is based on the notion of *incentivization* and on the claim that institutional facts follow from the fact that some institutional objects (*e.g.* a signpost, a traffic light, a piece of paper, …) are tied to a set of actions the individuals are incentivized to adopt. It is important to note that Smit et al. (2011, 7) do not see institutional objects to be the source of the incentives on the basis of which individuals act, but "they simply are *natural objects individuated by* these actions and incentives" (emphasis in original). A fuller definition of the notion of incentivization is the following:

> "Our basic view of how the logic of incentivization works is very simple. An actor (or actors) act(s) in such a way that another (or actors) is (are) incentivized to act in a specific way. The actions of the incentiviszing actors and the incentivized actions that result will differ for each social fact, but the deep logic is the same for all such facts. When such incentivization is stable enough the result is the kind of pattern of activity that we call a 'social fact'. When this process is directed at treating an *object* in a specific way the result is a Searlean 'institutional fact'" (Smit et al. 2011, 8, emphasis in original).

The authors discuss several examples to illustrate their notion of institutions as the product of an incentivization mechanism. The simplest one is the case of traffic lights (Smit et al. 2011, 5, emphasis in original): "A traffic light is only a traffic light because a set of actions are essentially tied to it. These actions include stopping on red, continuing to drive on green, and proceeding with caution on yellow. These actions are tied to traffic lights (…) in virtue of the fact that a given subject *S*, has been *incentivized* to act in the prescribed manner when encountering a traffic light". This example is straightforward: what makes a traffic light an institution (or an institutional fact) is the fact that individuals act on a basis of a set of incentives in such a way that it generates a distinctive or salient behavioral pattern relatively to a given object, in this case a traffic light. More complex examples follow exactly the same idea. For instance, what we call borders (one of Searle's favored examples) are nothing but a particular behavioral pattern consisting in the fact that some people cannot cross a line without being explicitly authorized to do so by some other people. This pattern is generated by some specific incentives which may have a variety of origins. Consider finally the institution of ownership (Smit et al. 2014): "A subject, *S*, owns an object, *o*, if, and only if, third parties are sufficiently disincentivized from interfering with *S*'s use of *o*". Once again, ownership as an institution is defined as a specific behavioral pattern, in this case a pattern of non-interference with one's use of an object. The latter is generated by some (not specified) incentives which lead each individual to act in such a way that the pattern is emerging.

The authors qualify their account as “naturalistic” or “reductionist”. Moreover, it follows from a “bottom-up” kind of explanation (Smit et al. 2014). By this, they mean that as far as ontology is concerned, we can account for the nature of an institution without making reference to any other institution. Only more “basic” concepts are needed, specifically the concepts of action and incentive: “On the incentivized action view it is claimed that institutional facts can be individuated, and understood completely, in terms of certain types of *actions*, namely the ones that we are incentivized to perform when confronted with some institutional object” (Smit et al. 2014). Moreover, the authors crucially point out that “the source of the incentives does not matter when conceptualizing the nature of the ‘institutional object’ itself” (Smit et al. 2011, 7). I take it to be Smit et al.’s more important but also more contentious claim regarding how to account for institutions. As I note in the preceding section, formally it translates as the claim that to know what lies behind the function $r_i(.)$ in expression (2) is irrelevant for a theory of institutions. What matters is that people are incentivized to act in a way which, at the aggregate level, generates a salient behavioral pattern relatively to some given object. The *reasons* motivating people to act in this way are irrelevant to qualify a behavioral pattern as an institution. There is one and only one criterion: that the individual decision rules r_i and the corresponding aggregative transition rule $\boldsymbol{r}$ create a dynamic with some fixed point $x^* = F(x^*)$, meaning that the agents are behaving in a self-enforcing manner.

Though Smit et al. ((2011); (2014)) do not explicitly formulate their incentivized action view in game-theoretic terms,[12] it is clear (as the above paragraph suggests) that it is nothing but an informal reformulation of the Standard Account. The concept of action is of course a centerpiece of game theory, even though we generally rather speak of “strategy”.[13] On the other side, the concept of incentives has no direct counterpart in a game-theoretic framework. However, incentives are tacitly encapsulated in the players’ preference ordering and belief function: if we assume that a rational player in a game maximizes his expected utility given his expectation, then the incentive to make a particular choice comes from what determines his preferences and expectation. As such, game theory is agnostic regarding the formation of the players’ preferences and beliefs. Hence, it could be considered that Smit et al.’s account is nothing but a restatement of a standard game-theoretic framework. I will return to this point below but for the moment it is sufficient to note that as long as game theory is considered as a “technology” to predict an outcome (or to prescribe a set of strategy recommendations) given a set of preferences and beliefs, the origin of preferences and beliefs and thus of incentives is indeed irrelevant.[14] Finally, the authors’ claim that “we can individuate institutional facts by the associated actions” (Smit et al. 2014) clearly indicates that institutions are nothing but patterns of actions. Of course, the very notion of “pattern” suggests some ergodicity in the

[12] Note however that in their discussion of the institution of money (Smit et al. 2011, 11), the authors explicitly note that their account basically consists of normal game-theoretical considerations. But this is obviously equally true for all of their other examples.

[13] Formally, a strategy is a complete list of contingent actions, *i.e.* decisions at information sets that may potentially be reached.

[14] There is a complication if payoffs and thus preferences are understood in a revealed preference perspective. In this case, preferences cannot explain the players’ choices since they are only a formal *description* of the latter. However, the claim made in the text is still true since we have no explanation for people’s choices.

history of plays which, as explained in the preceding paragraph, is equivalent to the notion of equilibrium in a game-theoretic framework.

On the basis of an analogy with the philosophy of mind literature, I shall argue that the conceptualization of institutions common to the Standard Account and Smit et al.'s incentivized action view is *eliminativist*. As said above, we could instead characterize them 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 both express a "summary view" of institutions. Indeed, as I show elsewhere (Hédoin 2014b), 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. However, it seems to me that the Standard Account and the incentivized action view are more properly qualified as eliminativist because they actually make the very concept of institution useless and irrelevant in the *explanation* of social phenomena. Consider the parallel of the philosophy of mind and the debate on the status on concepts referring to mental states or events: eliminativism (or "eliminative materialism") "is the view that certain common-sense mental states, such as beliefs and desires, do not exist" (Ramsey 2013). Basically, proponents of eliminativism in the philosophy of mind argue that the concepts of folk psychology are nothing but linguistic shortcuts to refer to only real events and states, *i.e.* events and states that occur at the neural level. Their point is the following: mental states referred to by the concepts of folk psychology do not really exist, and therefore "there is nothing more to the mind than what occurs in the brain" (Ramsey 2013). Eliminativists are also ontologically radical: they do not support the reduction or the displacement of traditional concepts referring to our folk ontology in favor of more fundamental ones; they argue more radically that these concepts are simply far too remote from reality to have any significance. They should be eliminated in favor of concepts referring to real entities. Mental states do not exist and we should stop referring to them, at least in philosophical and scientific discussions.

My point here is of course not to assess the relevance of eliminativism in the philosophy of mind. More relevant is the similarity between the eliminativism's underlying ontology and the accounts of institutions presented above. At first sight, to qualify the Standard Account and the incentivized action view as eliminativist may seem excessive. Smit et al. do not intend to eliminate the concept of institution; on the contrary they argue for a new way to understand it. Game theorists routinely speak of institutions and of related concepts such as conventions, norms or rules. Rather than arguing for the elimination of the concept, they contend that game theory is a highly useful tool to account for their emergence and their nature. However, at the explanatory level, the reduction of the concept of institution to more "basic" concepts such as actions/strategies, incentives, beliefs, and so on, ultimately results in its elimination. To paraphrase Ramsey (2013) on the relationship between the mind and the brain, there is nothing more to institutions than what occurs at the level of actions and incentives. Or, to say things otherwise: we do not need the concept of institution to explain social phenomena, but only concepts of actions, incentives, beliefs…

One significant consequence is that institutions are no longer distinctive properties of *human* societies: if institutions are nothing more than behavioral 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 confirm 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 aim in the next two sections is to propose a way to do so in a game-theoretic framework on the basis of a theory of rule-following.

4. Game Theory and Rule-Following Behavior

The thread of the two preceding sections has been that what game theorists are claiming to account for under the name "institution" are actually self-enforcing behavioral patterns that are the product of undetermined reasons for action. The rest of 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 behavior (*i.e.* a component of the *explanans*). In the latter case, the eliminativism of the Standard Account is necessarily wrong-headed because obviously we need the concept of institution as part of the explanatory endeavor. 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.

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.[15] 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

[15] 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 section and the next one can be seen as an attempt to go further in the formalization of these connections.

in following *R* (Hédoin 2014b). If we accept the idea that some institutions can be regarded as 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. In the formal framework of section 2, the constitutive rule enters into the function r_i which itself is responsible for the behavioral pattern the game-theorists call "institutions". 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 behavior 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 behavior of buyers and sellers without taking into account the rules that define a market exchange.

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 behavior (*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 behavior 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 behavior 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 theory to follow coordination rules (Lahno 2007, 446).

However, these judgments about the possibility to account for rule-following behavior in a game-theoretic framework seem too negative. Starting from Wittgenstein's writings and on the basis of David Lewis' theory of conventions (Lewis 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' 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 behavioral pattern to which it is ultimately reducible. However, as he notes himself, an agent may infer from the same behavioral 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.[16]

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' writings on conventions provide the right starting point. More particularly, it is Lewis' 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 usesthe 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, then he also has reason to believe that *P* holds (Lewis 1969, 52-53). 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*. There are many ways this insight can be incorporated in a game-theoretic framework. Here I propose to start from Robert Aumann's (1976) set-theoretic treatment of common knowledge because it is a standard way to treat issues related to knowledge and beliefs in a game-theoretic framework. However, as I will emphasize below, we will have to go beyond the pure set-theoretic framework to properly account for the fact that rule-following is the fact of using a mode of *reasoning* on the basis of the knowledge that it is shared in a population.[17]

5. Rule-Following in a Game-Theoretic and Epistemic Framework

We start from a game *G* described by the tuple $< N, S, \phi, \{u_i\}_{i \in N} >$ with the standard assumption that the *n* players are Bayesian rational, *i.e.* they maximize their expected utility on the basis of their beliefs and the latter are updated when new information is available according to Bayes' rule. The point is to relate the players' behavior 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 their knowledge of a particular rule*. To do this, it is necessary to

[16] 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 ontology. My understanding of intentional states in this paper is broadly 'externalist' in the sense defended by Don Ross ((2005); (2014)) on the basis of Daniel 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 'infer B from A', we are taking the intentional stance, *i.e.* an epistemological posture that helps us to make sense of the agent's behavior. In this sense, any intentional state is the product of a complex interaction the agent's behavior, the institutional context this behavior is embedded in, and the analyst's scientific goal. See Ross (2009) for the claim that Wittgenstein must be read as an externalist.

[17] Cubitt and Sugden (2003) have strongly argued against the use of Aumann's set-theoretic account to formalize Lewis' theory of common knowledge and particularly his concept of indication. Their argument is convincingly supported in particular by the fact that in a set-theoretic framework, all relations of implication are necessarily logical implications. Obviously, this does not capture what Lewis had in mind. I do not dispute this fact and it is precisely for this reason that we have to go beyond a pure set-theoretic treatment.

complete the description of G with an *information structure* $\mathscr{I}$. The resulting tuple $< N, S, \phi, \{u_i\}_{i\in N}, \mathscr{I}>$ describes what can be called an *epistemic game* $\mathscr{G}$. The information structure $\mathscr{I}$ has several components:

- A state space Ω where each state of the world $\omega \in \Omega$ is a complete description of everything that is relevant for the players, including their strategy choice, their beliefs and their knowledge.
- A possibility operator $\mathbf{P}_i$ for each $i \in N$ that defines an information (or knowledge) partition P_i over the state space Ω. A possibility operator indicates which state ω' a player i knows is possible when the actual state is ω. Formally, all such possible states are part of the set $\mathbf{P}_i\omega$.
- A prior belief $p_i(. ; \omega)$ defined over Ω for each $i \in N$.

The information structure $\mathscr{I}$ is thus a complete description of what each player knows and believes in each possible state of the world. The epistemic game $\mathscr{G}$ 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 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.

To fully articulate the notion of knowledge in an epistemic game, we furthermore need to define a *knowledge operator* **K** on the basis of the possibility operator **P**. First, we define an event E as a subset of Ω. An event can be understood as a set of states of the world which have at least one common feature (*e.g.* all states where it is raining outside). The event that an agent i knows that an event E happens (or has happened) is written $\mathbf{K}_i$E. Formally, an agent i knows that an event E holds at ω if and only if every state ω' he knows is possible, is a member of E, *i.e.* $\mathbf{K}_i\text{E} = \{\omega | \ \mathbf{P}_i\omega \subseteq \text{E}\}$. I will make the standard assumption throughout the section that the knowledge operator satisfies the axioms of the modal logic S-5 (Gintis 2009, 114):

For any events E and F,

(K1) $\mathbf{K}_i\Omega = \Omega$

(K2) $\mathbf{K}_i(\text{E} \cap \text{F}) = \mathbf{K}_i\text{E} \cap \mathbf{K}_i\text{F}$

(K3) $\mathbf{K}_i\text{E} \subseteq \text{E}$

(K4) $\mathbf{K}_i\text{E} = \mathbf{K}_i(\mathbf{K}_i\text{E})$

(K5) $\neg\mathbf{K}_i(\neg\mathbf{K}_i\text{E}) \subseteq \mathbf{K}_i\text{E}$

I will not comment on these five axioms in detail since they have already been largely discussed by game theorists.[18] I shall note two points however. Firstly, these axioms can be

[18] For instance, in addition to Gintis (2009), see Binmore and Brandenburger (1988).

deduced from the possibility operator **P** provided the latter satisfies two properties (Gintis 2009, 84):

(P1) $\omega \in \mathbf{P}_i\omega$
(P2) $\omega' \in \mathbf{P}_i\omega$ implies $\mathbf{P}_i\omega' = \mathbf{P}_i\omega$,
for any $\omega, \omega' \in \Omega$.

Property P1 states that an agent always knows that the current state is possible. Property P2 is required for P_i to be a partition: an agent cannot discriminate between two states that are part of the same partition. In other words, one's knowledge in two states that are in the same cell of one's information partitions must be identical. Secondly, while the first four axioms are mostly unproblematic,[19] the fifth one is largely unintuitive and demanding. It says that when *i* does not know something, he knows that he does not know this something. Though it is not required here strictly speaking, I will assume that it holds since it makes the exposition far easier.

We are now ready to return to Lewis' notion of indication and to the more general issue of rule-following behavior. I will build my account of rule-following in a game-theoretic framework by assuming that rule-following has two dimensions. Firslyt, a dimension of *reasoning*: to follow a rule is a fact about the way the players infer from an event a practical conclusion regarding what they should do. It is here that Lewis' notion of indication is relevant. Secondly, a dimension of *knowledge*: to follow a rule implies that the player knows the rule 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 the rule is *common knowledge* in the relevant population. My main point will thus be the following: in the case of rule-following, what must be common knowledge is not only each player's strategy and expectation; common knowledge of the players' reasoning is also required. When this latter condition holds, I will say that the players have a *common understanding of the situation* (Hédoin 2014a). I shall argue that such a common understanding is constitutive of rule-following behavior and is formally identical to Wittgenstein's concept of *lebensform*.

First we need to define a precise formulation of the concepts of *indication* and *common knowledge* in the epistemic framework set up above. I start with common knowledge since it is straightforwardly defined as an extension of the concept of knowledge in an epistemic game. Informally, a proposition *x* is common knowledge if each person knows *x*, each person knows that each person knows *x*, each person knows this, and so on. Formally, we can define common knowledge on the basis of either the knowledge operator **K** or the possibility operator **P**. In the latter case, a common knowledge event E is an event whose components are located at the *meet* of the players' information partitions P_i. The meet is defined by the communal possibility operator $\mathbf{P}^*\omega = \cup_{i \in N} \mathbf{P}_i\omega$, *i.e.* the union of the players' information partitions at ω. Since $\mathbf{P}^*\omega$ is the smallest set of states that all players know are possible at ω,

[19] K1 says that an agent knows the entire state space. K2 implies that an agent knows the logical implications of an event he knows (see below). K3 indicates that what an agent knows is always true (if one relaxes K3, we are no longer modeling knowledge but belief). Finally, K4 states that one knows what one knows. Note that all four axioms could be discussed and rejected on some ground.

an event E is obviously common knowledge if and only if $\mathbf{P^*}\omega \subseteq$ E, for all $\omega \in$ E. Then we are able to derive the common knowledge operator **K*** from **P*** in the same way we derived **K** from **P** above. Hence, we have $\mathbf{K^*}E = \{\omega|\mathbf{P^*}\omega \subseteq E\}$, meaning that at every state ω corresponding to E, everyone knows E, but also that everyone knows that everyone knows E, and so on. Such an event is sometimes called a *public event* because when E holds, it is transparent to everyone that E is common knowledge. Moreover, it should noted that the common knowledge operator **K*** also satisfies the five axioms of the S-5 modal logic. That implies that when E is common knowledge, this is necessarily common knowledge (K4) but also that if it is not common knowledge that E is not common knowledge, then E is common knowledge (K5). In other words, when an event is not common knowledge, this fact is commonly known among the players.

For the concept of indication, we now turn to Lewis' original statement. Lewis (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 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 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' reasoning in the epistemic and set-theoretic framework developed above.[20] I will only slightly change terms by speaking of knowledge instead of "reason to believe" and defining states of affairs as events. Then Lewis' three conditions can be reformulated as follows (see also Hédoin (2014a); Rescorla (2011); Sillari (2005); Vanderschraaf (1998)):

For all players $i, j \in N$, any events E and F, and any state $\omega \in \Omega$, if:
(L.1) $\omega \in \mathbf{K}_i E$
(L.2) $\mathbf{K}_i E \subseteq \mathbf{K}_i(\mathbf{K}_j E)$
(L.3) $\mathbf{K}_i E \subseteq \mathbf{K}_i F$
then $\omega \in \mathbf{K^*}F$.

[20] This statement needs to be qualified. See footnote 16 above.

The proof of this proposition is straightforward once we realize that the set-theoretic framework leads us to make several tacit assumptions regarding the way the agents are reasoning. Several things should be noted. Firstly, Lewis' notion of indication is here reduced to the logical implication "$\subseteq$". In particular, this framework does not allow us to distinguish deductive forms of reasoning from inductive ones. This is significant because, as I have already noted, rule-following is mainly about inductive reasoning. Secondly, in a set-theoretic framework, an event E indicates (or implies) an event F if and only if $E \subseteq F$ since if $\omega \in E$, then we also have $\omega \in F$. Then one must note that axiom K2 implies that whenever $E \subseteq F$ and $\mathbf{K}_i E$, then $\mathbf{K}_i E \subseteq \mathbf{K}_i F$ for every person *i*, which is condition L.3.[21] This result is essential: it states that an agent always knows the logical implications of what he knows. Thirdly, since the preceding result is a logical necessity, conditions L.1 to L.3 imply that the event F is common knowledge. Indeed, if E implies F, not only each agent *i* must know this; but each agent must also know that others know this and each agent must also know that others know that each know this. This may seem to be a surprising conclusion but actually it follows from the very nature of the set-theoretic framework, as emphasized by Aumann (1987). In such a framework, it is a logical truth that agents know the logical implications of what they know. Therefore, this must be true in all states ω. The reasoning abilities of the players are thus common knowledge in a casual sense. It follows from this that at any ω, each player is able to replicate the reasoning of the other players. Hence, if E implies F and if E is common knowledge, then F must be common knowledge.[22]

One way to capture this fact is to make the assumption explicit that the players are *symmetric reasoners* ((Gintis 2009); (Vanderschraaf 1998)) with respect to any event E. This corresponds to Lewis' ancillary conditions about "our rationality, inductive standards and background information". In a set-theoretic framework, these conditions are necessarily obtained because in an informal way they are implied by the construction of a state space Ω which is by definition commonly known among the players. This state space *is a constitutive part* of the agents' shared reasoning standards and background information. The condition of symmetric reasoning can then be stated in the following way (Gintis 2009, 142):

For any events E and F, any two players *i* and *j, i* is a symmetric reasoner with respect to *j* for E if

$$\text{(L.4) } (\mathbf{K}_i E \wedge \mathbf{K}_i(\mathbf{K}_j E)) \subseteq \mathbf{K}_i(\mathbf{K}_j F)$$

For the rest of this section, I will take for granted that conditions L.1-L.4 hold. Then, following Lewis (1969, 56), I will say that E is a basis for common knowledge of 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.

[21] The proof is straightforward: since $\omega \in \mathbf{K}_i E$ implies that $\mathbf{P}_i\omega \subseteq E$, it follows from $E \subseteq F$ that $\mathbf{P}_i\omega \subseteq F$. Hence, we also have $\omega \in \mathbf{K}_i F$, which in turn implies $\mathbf{K}_i E \subseteq \mathbf{K}_i F$.

[22] The proof is left to the reader. It parallels the proof of the preceding footnote, but by substituting the communal possibility operator $\mathbf{P^*}$ for the possibility operator $\mathbf{P}_i$.

Now, consider a game situation described by an epistemic game $\mathscr{G}$. What does it take for the players in $\mathscr{G}$ to follow a rule R? The first condition, following Lewis, is that the rule is common knowledge. Denote $R = \Omega$ as the event "rule R holds". Since the event corresponds to the entire state space, it is necessarily common knowledge. Hence, we can write $\mathbf{K^*}R$. However, while this event indicates itself, it does not as such give the players any hint regarding what to do. In the game $\mathscr{G}$, the *content* of the rule is specified by the players' priors p_i over the state space. For any given ω, each player i will settle on a strategy $s_i(\omega)$ on the basis of a conjecture $\sigma_i(\omega)$ that is determined by his prior p_i and his information partitions P_i, subject to the maximization condition

(7) $s_i \in \max_{s_i}\mathbf{E}[u_i(s_i\ ;\ \sigma_i)]$

Assume that the players' priors are heterogeneous, *i.e.* they disagree on how to play the game. In this case, they must hold inconsistent beliefs and their play will not result in an equilibrium. As a consequence, in spite of the fact that the rule is common knowledge, it fails to instruct the players a consistent way to play the game, *i.e.* the players are not symmetric reasoners with respect to R. Therefore, rule-following in a game $\mathscr{G}$ must depend on the fact that the players have a common prior $p(.\ ;\ \omega)$ over Ω. Indeed, Lewis' symmetric reasoning with respect to an event E implies that the players have a common prior over E (Hédoin 2014a). Now, we can properly define rule-following in $\mathscr{G}$:

> **Rule-following in $\mathscr{G}$**: A game $\mathscr{G}$ is *rule-governed* by a rule $R(\mathscr{G})$ whenever the players implement a strategy profile $\boldsymbol{s}^R(\omega) = (s_1(\omega), \ldots, s_n(\omega))$ on the basis of a set of conjectures $\sigma^R(\omega) = (\sigma_1(\omega), \ldots, \sigma_n(\omega))$ and where the following conditions hold:
>
> 1) The event $R = \Omega$ is common knowledge.
> 2) The n players share a common prior $p(.\ ;\ \omega)$ over Ω.
> 3) $\mathbf{E}[u_i(s_i\ ;\ \sigma_i)|\omega] \geq \mathbf{E}[u_i(\boldsymbol{s}^R_{-i}, g_i\ ;\ \sigma_i)|\omega]$, for all $i \in N$ and all ω, and where $\mathbf{E}$ is the expectation operator and g_i any other play by i than s_i.
>
> Then, R is reflexive common indicator for the strategy and belief profiles $\boldsymbol{s}^R(\omega)$ and $\sigma^R(\omega)$.

Together with the assumption that the players are Bayesian rational, the common prior assumption has an interesting implication: the strategy profile $\boldsymbol{s}^R(\omega)$ is a *correlated equilibrium* in $\mathscr{G}$ ((Aumann 1987); (Gintis 2009)). That means that the rule $R(\mathscr{G})$ assigns to each player i and for each state $\omega \in \Omega$ a strategy prescription on the basis of a function $r_i: \Omega \rightarrow S$ and the associated probability space (Ω, p). The function r_i and the probability space (Ω, p) together form a *correlated strategy*. Formally, the strategy profile $\boldsymbol{s}^R(\omega)$ then corresponds to the correlated distribution $\boldsymbol{r}(\omega) = (s_1 = r_1(\omega), \ldots, s_n = r_n(\omega); (\Omega, p))$ where condition 3 above is satisfied.

It should be noted that though I am defining a rule as a correlated equilibrium, this is not a restatement of the Standard Account. An institution refers to the whole phenomenon of rule-following behavior. The latter does not reduce to the behavioral pattern described by $\boldsymbol{s}^R$, but encompasses the whole practical and epistemic reasoning $\mathbf{K}^*R \rightarrow [\boldsymbol{s}^R(\omega) \wedge \sigma_P(\omega)]$ leading each agent to infer from the public event that a rule holds the strategy and belief profiles corresponding to the correlated equilibrium. To reason according to the practical inference $\mathbf{K}^*R \rightarrow [\boldsymbol{s}^R(\omega) \wedge \sigma_P(\omega)]$ *is* the fact of following the rule R. In this sense, the rule is constitutive of the whole practice ultimately leading to the self-enforcing pattern of play corresponding to the correlated equilibrium.

6. Rule-Following, Community Membership and Common Understanding

The game-theoretic account of rule-following developed in the preceding section clearly offers a distinctive view of institutions relatively to the Standard Account. However, it is conceptually very similar to Hurwicz's (1996) mechanism design approach. As I have noted in section 2, in the mechanism design literature, institutions are identified as game forms: they are the enforceable rules that determine (in the logical sense) the players' behavior. Similarly, the rule-following account of institutions I develop assimilates an institution to an entire epistemic game $\mathscr{G}$: $< N, S, \phi, \{u_i\}_{i \in N}, \mathscr{I}>$. An institution is thus not reduced to a behavioral pattern corresponding to an equilibrium play. It instead refers to the whole structure (including the information structure) in which the pattern of play is embedded. This account has the interesting property of making the fact that institutions (and hence, rule-following behavior) are constitutive of the players' reasoning and knowledge in a strategic interaction explicit: players reason as they do and know what they know *because* of a particular institution. Put differently, explaining individuals' behavior in a strategic interaction by referring to their reasoning and knowledge is the same as to say that institutions explain behavior. In this sense, institutions are truly constitutive rules.

Two related issues remain to be elucidated. First, if an institution is an epistemic game with a self-enforcing behavioral pattern, how does one interpret the fact that the players must have a common prior over the state space or, more generally, are symmetric reasoners with respect to the whole game? Second, what is the basis for this fact, *i.e.* on what conditions is the common prior assumption a reasonable explanation for a salient behavioral pattern? The first question is directly related to Lewis' "ancillary conditions" about the agents' rationality, inductive standards and background information. As I have explained in the preceding section, an agent i is a symmetric reasoner relatively to an agent j with respect to an event E if, when i knows E and knows that j knows E, then he knows that j infers the same conclusion than him. Symmetric reasoning is basically identical to the common prior assumption (Hédoin 2014a) since the latter guarantees that, provided the agents are endowed with the same information, they will form have the same posterior belief.[23] Therefore, given a public event E, players

[23] See Morris (1995) for a thorough and insightful examination of the common prior assumption in economics. The postulate that disagreement between rational agents come exclusively from differential information is called

with a common prior will necessarily infer the same conjecture regarding how each one will play in the game and this will also be common knowledge. The latter point may not be obvious but is a logical implication of the set-theoretic framework in which the players' belief and knowledge are modeled. Indeed, as Aumann (1987) points out, by construction the players' information partitions and prior belief are common knowledge: since each state is a complete description of everything that is relevant for the players, at each state all players know the others' information partitions and prior and this must be common knowledge in an informal sense (*i.e.* not defined by the knowledge operator $\mathbf{K^*}$ which is about event). Technically speaking, symmetric reasoning refers to the fact that properties of extensional equivalence between propositions are common knowledge.[24] This assumption is a structural property of set-theoretic models of knowledge and beliefs (Cubitt and Sugden 2003, 206-207).

Therefore, the set-theoretic framework in which we have modeled beliefs and knowledge implies some substantive assumptions regarding how agents reason and about what they know about each other agents reason. These assumptions (which are well beyond the scope of Bayesian rationality) are summarized in Hédoin (2014a) 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. Because common understanding necessarily holds in an epistemic game where the players have a common prior, 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 behavior a similar idea expressed under the concept of *lebensform* ("form of life") which more or less corresponds to an agreement in judgments between some persons (Wittgenstein 1953, §242). As Sillari (2013, 884-5) rightly notes, such an agreement has a natural counterpart in Lewis' 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.

The second issue follows almost immediately: 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 (Wiitgenstein 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

"Harsanyi's doctrine" by Aumann (1987). As Aumann notes, Harsanyi's doctrine is basically the common prior assumption.

[24] For a state ω, consider the events $\mathbf{K}_i\mathrm{E}$ and $\mathbf{K}_j(\mathbf{K}_i\mathrm{E})$ with $\mathbf{K}_i\mathrm{E}$ and $\mathbf{K}_j(\mathbf{K}_i\mathrm{E})$. The definition of a state (and the whole definition of a state space) then implies that the expression "*i* knows that E holds" is extensionally equivalent to the expression "*j* knows that E' holds" for some event E' where *j* knows that E' holds and where E' contains all those states where *i* knows that E holds.

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 behavioral 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 skeptical 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 1953, §85).

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 establish custom or usage (Wittgenstein 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 (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 know the rule. In other words, what makes two persons belong to the same community is the fact that they follow the same rule(s), and thus that they share a form of life, *i.e.* a set inductive standards. In the game-theoretic framework of the preceding section, a game situation depicted by the epistemic game $\mathcal{G}$ where the players have a common prior $p(. ; \omega)$ makes the latter the members of a community in which, by definition, they follow a rule which gives rise to the behavioral pattern corresponding to a correlated equilibrium. Community membership thus implies common understanding and thus rule-following but at the same time, rule-following behavior is constitutive of community membership. At this point, the account of institutions as epistemic games and Wittgenstein's account of rule-following totally converge.

7. 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 definition of institutions as epistemic games that significantly departs from the standard game-theoretic account of institutions, where the latter are defined as mere behavioral regularities.

I would like to finish this article by briefly considering two objections that could be formulated against my account. Firstly, if the charge against the eliminativist view developed in the first sections 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 behavioral regularities should not be called "institutions" or that the "essence" of institutions is something different. This sort of metaphysical considerations has little value from a scientific, if not philosophical point of view. Rather, what have 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 behavior. 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.

Secondly, the definition of institutions as epistemic games may be rejected on two substantive grounds: on the one hand, it relies on unnecessarily and unrealistically strong epistemic requirements, in particular the fact that institutions depend on public events together with the common prior assumption; on the other hand, this definition is unhelpful as far as the issue of the emergence and the evolution of institutions is at stake. The first critique is developed for instance by Binmore (2008) who argues that public events in the social world are probably rare if not non-existent. The common prior assumption is also very strong because it implies that persons agree regarding the probabilities of non-natural events. Though these arguments are of some value, 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 humans' ability to form recursive chains of knowledge (eventually leading to common knowledge) is what distinguishes humans from other animals. Finally, the common prior assumption is itself justified by the fact that institutions imply that people have a common understanding of the situation or, in Wittgensteinian terms, agree over a *lebensform*. The second critique is well beyond the scope of this paper. A short answer has already been suggested in the preceding paragraph: the point is not to reject the Standard Account of institutions which indeed, provides a bottom-up explanation of the emergence of behavioral patterns. The definition of institutions as epistemic games emphasizes a distinct aspect of the social world – the cognitive and epistemic mechanisms lying behind coordination and cooperation. In fact, as Aoki (2001) has already suggested, both approaches may be highly complementary.

References

Aoki, Masahiko. 2001. *Toward a Comparative Institutional Analysis*. MIT Press.

Aumann, Robert J. 1976. "Agreeing to Disagree." *The Annals of Statistics* 4(6): 1236–39.

———. 1987. "Correlated Equilibrium as an Expression of Bayesian Rationality." *Econometrica* 55(1): 1–18.

Aumann, Robert J., and Jacques H. Dreze. 2008. “Rational Expectations in Games.” *The American Economic Review* 98(1): 72–86.

Binmore, Ken. 2008. “Do Conventions Need to Be Common Knowledge?” *Topoi* 27(1): 17–27.

Binmore, Ken, and Adam Brandenburger. 1988. “Common Knoweldge and Game Theory.” *CREST Working Paper* 89-06, University of Michigan.

Bloor, David. 1997. *Wittgenstein, Rules and Institutions*. Routledge.

Chwe, Michael Suk-Young. 2003. *Rational Ritual: Culture, Coordination, and Common Knowledge*. Princeton University Press.

Cubitt, Robin P., and Robert Sugden. 2003. “Common Knowledge, Salience and Convention: A Reconstruction of David Lewis’ Game Theory.” *Economics and Philosophy* 19(2): 175–210.

Fudenberg, Drew, and David K. Levine. 1998. *The Theory of Learning in Games*. Cambridge, Mass: The MIT Press.

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

Greif, Avner. 2006. *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade*. Cambridge University Press.

Greif, Avner, Paul Milgrom, and Barry R. Weingast. 1994. “Coordination, Commitment, and Enforcement: The Case of the Merchant Guild.” *Journal of Political Economy* 102(4): 745–76.

Hausman, Daniel M. 2011. *Preferences, Value, Choice, and Welfare*. Cambridge University Press.

Hédoin, Cyril. 2014a. “A Framework for Community-Based Salience: Common Knowledge, Common Understanding and Community Membership”, *Economics and Philosophy* 30(3): 365-395.

Hédoin, Cyril. 2014b. “Accounting for Constitutive Rules in Game Theory”, *Journal of Economic Methodology*, forthcoming.

Hodgson, Geoffrey. 2006. “What Are Institutions?”, *Journal of Economic Issues* 40(1): 1-25.

Hurwicz, Leonid. 1996. “Institutions as Families of Game Forms.” *Japanese Economic Review* 47(2): 113–32.

Lahno, Bernd. 2007. “Rational Choice and Rule-Following Behavior.” *Rationality and Society* 19(4): 425–50.

Lewis, David K. 2002. *Convention: A Philosophical Study*. John Wiley and Sons.

Maynard Smith, John. 1982. *Evolution and the Theory of Games*. Cambridge University Press.

Milgrom, Paul R., Douglass C. North, and Barry R. Weingast. 1990. “The Role of Institutions in the Revival of Trade: The Law Merchant, Private Judges, and the Champagne Fairs.” *Economics & Politics* 2(1): 1–23.

Morris, Stephen. 1995. “The Common Prior Assumption in Economic Theory.” *Economics and Philosophy* 11(2): 227–53.

North, Douglass C. 1990. *Institutions, Institutional Change and Economic Performance*. Cambridge University Press.

Ramsey, William. 2013. "Eliminative Materialism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2013.

Rawls, John. 1955. "Two Concepts of Rules." *The Philosophical Review* 64(1).

Rescorla, Michael. 2011. "Convention." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2011.

Ross, Don. 2005. *Economic Theory and Cognitive Science. Microexplanation*. The MIT Press.

Ross, Don. 2009. "Reply to Hands: On the Robbins-Samuelson Argument Pattern." *Journal of the History of Economic Thought* 31(1): 93-103.

Ross, Don. 2014. *Philosophy of Economics*. Palgrave Macmillan.

Schotter, Andrew. 2008. *The Economic Theory of Social Institutions*. Cambridge University Press.

Searle, John R. 1995. *The Construction of Social Reality*. Simon and Schuster.

———. 2010. *Making the Social World: The Structure of Human Civilization*. Oxford University Press.

Sillari, Giacomo. 2005. "A Logical Framework for Convention." *Synthese* 147(2): 379–400.

———. 2013. "Rule-Following as Coordination: A Game-Theoretic Approach." *Synthese* 190(5): 871–90.

Skyrms, Brian. 1996. *Evolution of the Social Contract*. Cambridge University Press.

Smit, J. P., Filip Buekens, and Stan du Plessis. 2011. "What Is Money? An Alternative to Searle's Institutional Facts." *Economics and Philosophy* 27(1): 1–22.

Smit, J. P., Filip Buekens, and Stan du Plessis. 2014. "Developing the Incentivized Action View of Institutional Reality." *Synthese*, forthcoming.

Sugden, Robert. 1991. "Rational Choice: A Survey of Contributions from Economics and Philosophy." *The Economic Journal* 101(407): 751–85.

———. 2005. *The Economics of Rights, Cooperation and Welfare*. 2nd ed. Palgrave Macmillan.

Tomasello, Michael. 2014. *A Natural History of Human Thinking*.

Vanberg, Viktor J. 2004. "The Rationality Postulate in Economics: Its Ambiguity, Its Deficiency and Its Evolutionary Alternative." *Journal of Economic Methodology* 11(1): 1–29.

Vanderschraaf, Peter. 1998. "Knowledge, Equilibrium and Convention." *Erkenntnis* 49(3): 337–69.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. John Wiley & Sons.