Article

Cyril Hédoin*

The Beliefs-Rules-Equilibrium Account of Institutions: A Contribution to a Naturalistic Social Ontology

https://doi.org/10.1515/jso-2020-0001
Published online April 23, 2021

Abstract: This paper pursues a naturalist endeavor in social ontology by arguing that the Beliefs-Rules-Equilibrium account of institutions can help to advance the debate over the nature of social kinds. This account of institutions emerges from a growing number of works in economics that use game theory to study the role and the functioning of institutions in human societies. I intend to show how recent developments in the economic analysis of rules and institutions can help solve issues that are generally considered constitutive of any ontological inquiry. I argue that the Beliefs-Rules-Equilibrium account of institutions can contribute to advancing the debate on an issue of particular importance, regarding the specific form of dependence characterizing the relation between institutions and individuals' attitudes about them. I tackle this issue by taking Francesco Guala's claims about the nature of institutions made in his book "Understanding Institutions" as a point of departure. In particular, I reject Guala's functionalism about institutions. On the basis of the Beliefs-Rules-Equilibrium account, I claim that it is futile to search for constitutive features of general institutions (money, property rights, family…) and that the best we can have is a knowledge of what are the rules within a specific institution, which the agents consider to be essential in their institutional practice.

Keywords: beliefs-rules-equilibrium account, institutions, realism, social ontology, social kinds

***Corresponding author: Cyril Hédoin**, Department of Economics, Economics and Management Research Center REGARDS, University of Reims Champagne-Ardenne, 57B Rue Pierre Taittinger, Reims, Marne, 51096, France, E-mail: cyril.hedoin@univ-reims.fr. https://orcid.org/0000-0002-5358-7111

1 Introduction

As the field concerned with the nature of the social world and of social objects, social ontology is by all means in close relationship with social sciences. Though perhaps schematic, two broad perspectives on the relationship between social ontology and social sciences can be distinguished on the basis of recent studies in this field. From a '*foundationalist*' perspective, social ontology is viewed as providing the metaphysical foundations on the basis of which sound scientific modeling and theorizing can be developed within the social sciences. Such a perspective is notably explicitly argued for by philosophers like John Searle (2010), Raimo Tuomela (2013) or Brian Epstein (2015). The foundationalist view clearly gives social ontology a conceptual priority, in the sense that an appropriate scientific understanding of the social world cannot be attained without first 'getting the ontology right'. On the other hand, the '*naturalist*' perspective follows the broader lead of naturalism in philosophy, by postulating that ontological knowledge and scientific knowledge are essentially of the same kind and can be produced along essentially similar methods. This view has notably been recently endorsed by economists and philosophers of economics, e.g. Guala (2016) or Smit et al. (2011). The naturalist perspective considers that social sciences determine the appropriate constraints that must be imposed to any relevant social ontology and discards the existence of social ontology as an independent field, arguing for 'foundationless social sciences' (Sugden 2016).

This paper is a contribution to the naturalist perspective. I intend to show how recent developments in the economic analysis of rules and institutions can help solve issues that are generally considered as constitutive of any ontological inquiry. More precisely, I argue that, what I call the *Beliefs-Rules-Equilibrium* account of institutions, which characterizes recent works in the economics of institutions, can contribute to answering several questions in social ontology about the nature of "social kinds". I focus on an issue of particular importance regarding the specific form of dependence that characterizes the relation between institutions and individuals' attitudes about them. I tackle this issue by taking Guala's (2016) claims about the nature of institutions as a point of departure. In particular, I reject Guala's functionalism about institutions. On the basis of the *Beliefs-Rules-Equilibrium* account, I claim that it is futile to search for constitutive features of general institutions (money, property rights, family…) and that the best we can have is a knowledge of what are the rules within a specific institution that the agents consider to be essential in their institutional practice. The overall contribution of the paper is thus twofold. First, it pursues the project of building a naturalistic social ontology, as initiated by Hindriks and Guala (2015) and Smit et al. (2011)

among others, by making more specific ontological claims. Second, it aims at advancing on this basis a new proposition regarding the nature of institutions and the way in which their essential rules are conceived.

The rest of the paper is organized as follows. Section 2 presents the *Beliefs-Rules-Equilibrium* account of institutions. Section 3 argues for a distinction between social kinds and natural kinds on the basis of the existence of an 'internal point of view' that can be taken to study the former but not the latter. Section 4 rejects Guala's functionalist approach to characterize institutions and presents an alternative view in terms of essential rules. Section 5 concludes by considering the prospects of a general theory of institutions.

2 The Beliefs-Rules-Equilibrium Account of Institutions

At the most general level, an institution is a set of rules, such that agents are incentivized to behave in a relatively regular and predictive way. This definition is neither conceptually nor theoretically innocuous: it is derived from a growing body of game-theoretic analyses of institutions developed by economists and other social scientists.[1] In spite of some heterogeneity, this body of works captures the constitutive features of, what I call, the Beliefs-Rules-Equilibrium (henceforth, BRE) account of institutions. This section provides a characterization of this account. Unsurprisingly, given the fact that it builds on a similar body of works, the BRE account shares many features of Hindriks and Guala's (2015) recent rules-in-equilibrium approach. I shall argue, however, that the BRE account is at the same time more general (i.e. applies to a wider range of cases) and makes more specific ontological claims for reasons explained below.

Hindriks and Guala (2015) distinguish two generic accounts of institutions within economics. On the institutions-as-equilibria account, institutions correspond to equilibria in games. More specifically, they are defined as stable (in some sense)

[1] It is not possible to cite the whole body of relevant literature here. Schelling (1981) and Sugden (2005) figure as two of the earliest and most significant game-theoretic accounts of institutions. Greif (2006) presents several methodologically related historical studies of particular economic institutions that build on a game-theoretic framework. Basu (2018) develops, what he refers to as, the 'focal point approach' within the economic analysis of law. This approach applies the more general game-theoretic account of institutions to law and economics, as sketched notably by Hurwicz (1996, 2008) and Myerson (2009). Other major significant contributions are by Bicchieri (2006), Binmore (1998), Schelling (1981) and Skyrms (1996).

patterns of behavior such that no agent has an incentive to change her behavior: “The defining characteristic of an equilibrium – what distinguishes it from other profiles – is that each strategy must be a best response to the action of the other players or, in other words, that no player has an incentive to change her strategy unilaterally. If the others do their part in the equilibrium, no player can do better by deviating” (Hindriks and Guala 2015: 6). On the institutions-as-rules account, institutions are rather identified as (set of) rules guiding agents’ conduct. Institutions are ‘the rules of the game’ indicating what is permitted, obligatory or forbidden. They are therefore thought to facilitate human interactions and to encourage activities, or quite the contrary, to discourage others: “institutional economists like North have used the rule conception to study the way in which institutions facilitate growth for example. Accountancy rules foster transparency and trust; bankruptcy rules reduce uncertainty when businesses fail; property rights encourage investments, and so forth” (Hindriks and Guala 2015: 4).

Each account has its shortcomings. The institutions-as-rules account misses the important fact that rules are effective only if agents are properly incentivized to follow them and lacks an adequate explanation of the mechanisms guaranteeing that such incentives exist. The institutions-as-equilibria account takes a mostly behavioral stance and ignores the fact that rules and institutions are not mere summaries of behavioral patterns but are also cognitive instruments that agents use to form appropriate intentional attitudes.[2] Hindriks and Guala (2015) and Guala (2016) argue for a “rules-in-equilibrium” account, unifying the institutions-as-rules and institutions-as-equilibria views, according to which institutions are *both* rules and equilibria. In this account, rules correspond to cognitive devices used by individuals to represent the equilibria they are playing. This is indeed an important idea that I propose to generalize and at the same time to make more specific through the BRE account of institutions.

The BRE account builds on the postulate that institutions are a set of rules that operate within what Basu (2018) calls the ‘game of life’.[3] The game of life is a game-theoretic model of everyday interactions in which humans are participating. It is a full description of all the strategies physically available to all agents, as well as of all the consequences that can be attached to any combination of these strategies. In other words, the game of life is just the *complete game-form* including all events

[2] The distinction between the conception of rules as summaries of behavior and the conception of rules as instruments used in the context of some practice is for instance made by Wittgenstein (1965) and Rawls (1955). I return on the significance of this distinction in the next section.

[3] The concept first appeared in Binmore (1998).

(strategies and outcomes) that are possible according to the laws of nature.[4] By assumption, the game of life is immutable: the sets of players, strategies and outcomes cannot be altered by any human interventions – except, maybe for significant and exceptional technological change.

How to account for the role and functioning of institutions within this game of life? The label BRE provides an indication: institutions (partially) determine agents' behavior on the basis of rules through which agents form beliefs about the social world (including the behavior of others); these beliefs in turn, combined with the agents' preferences over outcomes, are incentivizing agents to behave in such a way that – providing their beliefs are approximately correct – no one wishes to change her behavior, i.e. the resulting strategy profile is a game-theoretic equilibrium. This formulation is an open one in the sense that it does not prescribe a specific way to model social institutions. In particular, it does not indicate which equilibrium concept is the relevant one. It is a feature that makes the BRE account a generalization of recent naturalistic social ontologies. This is in particular the most important difference with Guala's (2016) and Hindriks and Guala's (2015) 'rules-in-equilibrium' account, which makes use of the correlated equilibrium solution concept. The substantive implication is that institutions are defined as correlation devices (see also Gintis 2009). In some cases, this may be unnecessarily restrictive, for instance when a rule fails to prescribe a definite behavior. Leaving of the issue of the relevant solution concept opened makes the BRE account more general, since it applies to a larger class of games, including dynamic games formalizing sequential interactions. At the same time, however, it provides some precise indications regarding the mechanisms that establish a relationship between the existence of an institution and the agents' behavior. On this basis, I shall characterize the BRE account of institutions through the following four points. While neither of these points contradicts Hindriks and Guala's (2015) rules-in-equilibrium account, each of them makes more specific claims regarding the nature of institutions that directly account for the importance of, what I shall refer to as, the internal point of view. In this sense, I contend that the BRE account not only generalizes Hindriks and Guala's, but that it also leads to more determinate answers to some of the ontological issues discussed below.

First, an institution restricts the set of available strategies for each player. It does not do so by making *physically* impossible for an agent to make some strategic

[4] Technically, a game form F is a triple $<N, S, O>$ where N is a set of players or agents, S is the set of pure strategies defining possible strategy choices to each player and O is the set of outcomes or consequences that result when any strategy profile s belonging to S is implemented. F is not a game properly speaking because the definition of a game requires to specify the players' preferences over O. See Myerson (2009) for a semi-technical discussion.

play, but rather by ensuring that the players' beliefs are such that they will not have an incentive to play literally 'against the rules'. For instance, in chess, while it is physically possible for a player to move her pawns backward, or even to punch her opponent in the face, such behavior is (almost) never observed during a chess game. Similarly, even though it may occur that a driver attempts to bribe a police officer to avoid a fine, this is not a behavior that is normally seen in many countries. In both cases, what happens is that the institution defines, what can be called, a *social game* in which the set of socially permissible strategies is a strict subset of the physically possible ones, with the set of possible outcomes correspondingly reduced.[5] Depending on the criterion used to demarcate socially permissible strategies from socially forbidden ones, it is of course possible that several social games may be figured out by different players in the very same circumstances. Thus, the very effectiveness of an institution builds on a form of (most likely tacit) agreement among the players that is formally captured by shared (or at least consistent) beliefs over what one *is not likely* to do.[6]

Second, the rules constituting an institution may sometimes work in a hierarchically-ordered way. In the simplest case, there may be primary (or first-order) rules socially forbidding some strategic play and secondary (or second-order) rules specifying – *given the social game actually played* – which strategic outcome should be implemented. If the primary rules socially forbid all but one strategy for each player, then of course there is no need for secondary rules. In a more complex case, primary rules may indicate how to determine the relevant social game to be played in specific circumstances. Secondary rules will then specify the socially impermissible strategic plays, while ternary rules will eventually indicate which outcome should be implemented, again given the social game actually played. In principle, an infinite hierarchy of such ordered rules may exist, but it is more relevant, both conceptually and practically, to assume that such a hierarchy must be finite. This is of course a common Wittgensteinian theme that the infinite regress of interpreting rules with rules must ultimately come to an end. In the philosophy of

[5] It might be argued that the bribery example is different from the chess example, because in the former compliance with the rule that forbids bribery is likely due to there being enforceable laws against such behavior. But the difference is only apparent because what makes the laws enforceable and actually enforced is a rule or set of rules excluding some actions (e.g. to not sanction the acceptance of a bribe) from the social game.

[6] The concept of social games I put forward is formally identical to Hurwicz's (2008) concept of 'legal game'. Hurwicz proposes a criterion of dominance to determine whether a legal game G is enforceable within the game of life F. Successful enforcement then requires that every illegal strategy is dominated in a game-theoretic sense by a legal strategy. A weaker criterion suggested by Myerson (2008) is that only legal strategies should be the best response for a player who expects everyone else to play a legal strategy. Technically, G then corresponds to a *curb set* in F.

law, Hart's (2012) distinction between 'rules of behavior', indicating how to behave, and 'rules of recognition', specifying criteria of legal validity, may be viewed as an exemplar of the more general view that institutions are systems of hierarchically-ordered rules. Within the BRE account, such view has at least two related implications. A first implication is that higher-order rules may be functionally dependent on lower-order ones. Consider for instance the case of marriage.[7] A first-order rule may authorize same-sex unions. Second-order rules may then state specific conditions regarding filiation in such unions, rules which presumably may differ from those applying for different-sex unions. However, should same-sex unions become proscribed, these second-order rules would obviously become irrelevant. This leads to a second implication, which I shall more thoroughly explore below, namely the fact that some rules may be regarded as more essential or fundamental than others in determining the nature of an institution. How a rule is located in the hierarchy constitutive of a given institution may be relevant in determining whether a rule should be regarded as essential or not, though being a first-order rule is neither necessary nor sufficient to be ascribed to such status. This is due to the fact that interfering with a lower-order rule may affect one or several higher-order rules, while the reverse is generally not the case.

Third, the very reason why rules are needed in the first place is that they help solve coordination and cooperation problems. More specifically, within any given social game G, several strategy profiles (i.e. combination of strategic plays) may correspond to an equilibrium. Moreover, some of them may be better than others, according to relevant normative criteria. Rules are then devices through which agents determine what they should do by forming expectations about what others will do. Similarly, given the game of life F, there may be several possible social games G that can potentially be played. Participants in a social game must then be able to determine, which social game is actually played and eventually what to do in it. In other words, players have to form the right set of beliefs over what others are doing and believing. Schelling's (1981) concept of *focal points* captures the relationship between rules and beliefs: a rule determines what an agent sees as 'self-evident' or 'the most likely', because it consists in a shared understanding of a given social interaction. The existence of focal points itself depends on shared inductive standards and reasoning modes (Hédoin 2014; Lewis 2002; Sugden 2011). In this sense, rules and beliefs are not causally related (e.g. a rule causally determines one's beliefs), but are rather constitutively related: the existence of a

[7] As I argue in section 5, one should be cautious working with concepts of general institutions, which may fail to fully capture any specific instance of a particular, spatially and temporally located institution. However, using such concepts is not a problem, as long as one does not ascribe normative significance to them.

rule *consists in* the fact that an outcome is identified as a focal point or a social game is identified as a 'focal curb'.[8] As noted by Basu (2018), this is a key insight for understanding how law can change people's behavior: a law never creates new strategies *de novo* or directly alters outcomes in the game of life. A law-induced change in behavior is always due to a successful change in people's beliefs, either in the game of life or in a social game. The case of law illustrates a more general point about the way a rule works in shaping individuals' beliefs and behavior. While social outcomes are obviously causally dependent on people's choices, the latter in turn depend on how people are reasoning and especially on the kinds of inferences they are making from (partially or fully) public events. 'Focal point' is actually a catchword for the fact that, within specific 'game-situations', all individuals are symmetrically reasoning, i.e. they are inferring the same conclusion from the same event (Gintis 2009). This is this 'meeting of minds' that makes particular outcomes look as self-evident.

Institutions are therefore viewed as hierarchical sets of rules indicating what is socially permissible (the social game defined as a focal curb) and eventually what is socially expected (the focal point within the social game). These rules are themselves related to a nexus of beliefs and choices: to claim that a game is rule-governed implies that the players share some inductive standards and reasoning modes, such that in a given 'game situation' all players share convergent expectations about what others are believing and doing, i.e. that a focal point exists. The very existence of an institution (or of an 'institutional kind') thus entails, on the BRE account that:

(1) Agents share inductive standards and reasoning modes, what corresponds to the existence of focal points.
(2) Agents hold correct beliefs about others' beliefs and choices, as implied by the preceding point.
(3) The choices form a game-theoretic equilibrium, such that everyone is incentivized in acting as they do.

Fourth, a rule not only prescribes what should actually be done or not in a given set of circumstances. It also indicates what *would* have to be done or not if the circumstances had been different. In other words, the existence of rules is deeply intertwined with the cognitive possibility of *conditional reasoning* and thus with the ability of agents to form *conditional beliefs* (Hédoin 2019).

[8] I am not claiming that the existence of an institution implies complete determinacy. Institutional practices may be partially indeterminate. For instance, a rule may determine which social game is played in a given set of circumstances but not what should be done in this game. In this case, while there is a focal curb, no focal point exists and the coordination is likely to be imperfect.

In particular, the functioning of institutions is partially determined by the players' counterfactual beliefs over null events, i.e. events to which they ascribe a zero probability. The importance of conditional reasoning over such 'impossible' events can be game-theoretically established for historically well-identified institutions.

Hédoin (2019) illustrates this point on the basis of Avner Greif's (2006: chap. 9) comparative study of the organization of economic exchanges in two communities of traders in the period of the Middle Ages: the Maghribi traders (descendants of Jewish traders who first emigrated to North Africa and then to Egypt) and the Genoese traders. These two communities were facing the same commitment problem regarding overseas trade: it was generally not possible for a trader to embark overseas to trade with local merchants in other countries. So Maghribi and Genoese merchants used to hire "agents" representing their interests abroad. Agents were paid a wage by merchants and had the responsibility to keep the merchandise safe and to negotiate exchange terms with local merchants. This is a classical principal–agent relationship, which poses the usual moral hazard problem. Greif shows that these two communities solved the commitment problem through two different sets of institutions. The most interesting feature of Greif's analysis is his argument that the institutional divergence between these two communities of traders is explained by the fact that their members held different "cultural beliefs". Cultural beliefs are "the shared ideas and thoughts that govern interactions among individuals and between them, their gods, and other groups". They "differ from knowledge in that they are not empirically discovered or analytically proved". Finally, they "become identical and commonly known through the socialization process, by which culture is unified, maintained, and communicated" (Greif 2006: 269–70). Cultural beliefs are directly responsible for the equilibrium selection in a game, because they provide focal points and help the coordination of expectations. They are self-enforcing, since at the equilibrium the players' beliefs are correct, i.e. they match with the actual behavioral pattern corresponding to the institutional practice. Greif explicitly characterizes self-enforcing cultural beliefs as "a set of probability distributions over an equilibrium strategy combination". In particular, each probability distribution "reflects the expectation of a player with respect to the actions that will be taken *on and off the path of play*" (Greif 2006: 270–1, my emphasis). In other words, cultural beliefs extend over events, which cannot occur under the rule-governed practice. Greif's study shows that these off-the-path-of-play beliefs are nonetheless decisive because they directly determine, within the related game-theoretic model, the threshold values of parameters or variables (in this case, wage level), making a given behavioral pattern (strategy profile) an equilibrium. Hence, a change in these beliefs would entail a change in the equilibrium and thus a different rule.

The importance of conditional reasoning is also related to what characterizes institutions as *social kinds*, i.e. the fact that they cannot be studied without acknowledging the existence of an *internal point of view of rules*. As I have explained, individuals following a rule have to consider counterfactuals about what they should do in situations that cannot occur if the rule is indeed followed. This indicates that rules are more than behavioral patterns. Below the behavioral surface, rules are followed for reasons that may not be behaviorally transparent. I shall argue that the existence of these reasons, captured in a normative language, are a distinctive characteristic of social kinds. This will pave the way for my argument against Guala's functionalism about institutions and in favor of an approach in terms of essential rules made in section 4.

3 Social Kinds, Normativity and the Internal Point of View of Rules

The concept of 'internal point of view of rules" is due to the legal philosopher Herbert L. A. Hart (2012). The broad idea – which can be traced back at least to the hermeneutic philosophy of the late nineteenth century – is that when studying an object entering into the category of social kinds, such as an institution or a social practice, one can take two different points of view. Statements made from the external point of view generally consist in describing some regularities of behavior within a population. They may eventually go further and make reference to rules that are assumed to guide and account for individuals' behavior. The reference to rules is, however, a mere shortcut to make probabilistic predictions about the behavior of individuals, since the observer is not taking any stance about the 'realisticness' of those rules. Actually, the external point of view is as relevant for studying animal societies as human societies. Ants or bees behave in a largely predictive way and these behavioral regularities can be regarded as corresponding to 'rules' easily analyzable as game-theoretic equilibria.

The external point of view does not exhaust, however, the perspectives that can be taken to analyze institutions and social practices. Consider someone observing that individuals are regularly and predictably stopping at red traffic lights. Adopting the external point of view would merely consist in stating that there is a correlation between the traffic light and the behavior within a population. However, by restricting themselves to the external point of view,

> [s]he will miss out a whole dimension of the social life of those whom [s]he is watching, since for them the red light is not merely a sign that others will stop: they look upon it as a *signal for* them to stop, and so a reason for stopping in conformity to rules which make stopping when the light is red a standard of behavior and an obligation. To mention this is to bring into the account the way in which the group regards its own behavior (Hart 2012: 90).

According to Hart, the internal point of view of rules is related to what he called their 'internal aspects' and which distinguishes them from mere social habits leading to behavioral regularities. The internal aspect refers to the fact that below the behavioral surface, rules are followed for specific *reasons* that are usually captured by the use of a normative language. In other words, there is a normativity inherent to the very existence of rules that escapes the external point of view. Figuring this normativity is not only important in some cases to explain individuals' behavior within a social practice. It is also essential to account for the fact that rules may serve as a *justification* for one's behavior.

A very similar idea has been put forward by John Rawls in his article "Two Concepts of Rules" (Rawls 1955). Rawls begins with the distinction between two kinds of justification: the justification of a practice on the one hand, and the justification of an action within a practice on the other hand. In many cases, the two kinds of justification will not build on the same set of principles. Consider for instance the practice of promise-keeping. While the practice itself may be plausibly defended on the ground of a utilitarian principle, there will be many particular instances where the very same principle shall recommend that one does not keep their promise. However, as soon as a particular action falls within the practice of promise-keeping, the reason for keeping one's promise no longer lies in the utilitarian principle but instead in the commitment that is created when one abides by the practice of promise-keeping. That does not mean that breaking one's promise is universally and unconditionally forbidden under the practice of promise-keeping, only that the normative ground for justifying so cannot be the same as the one justifying the practice as a whole.

Taking this distinction for granted, Rawls argues that another distinction should be made, this time between the two concepts of rules: the *summary view* and the *practice conception*. The former conceives rules in terms of statistical summaries of past behaviors resulting from the application of some normative principle of decision-making (e.g. the utilitarian principle). The latter views rules as being logically prior to any instantiation of a practice. Rather than summarizing past behavior, rules create the conditions for making some actions *logically* possible. Rawls illustrates this point by taking the example of the game of baseball,[9] but it is easy to extend it to socioeconomic institutions too. To marry someone logically presupposes a rule defining what it is to be married. Making counterfeit money (and eventually being punished for that) is logically impossible without a set of rules defining what money and counterfeit money are. Stealing someone's property (and eventually being punished for that) is logically dependent on a concept of property corresponding to a set of rules defining it. The relevance of this distinction comes from the fact that, according to Rawls, only the practice conception is able to account for the difference between the two kinds of justification he introduces. This is especially salient with respect to the authority of rule-following, which is precisely the issue Hart considers through his account of the internal aspect of rules. As Rawls (1955: 26) puts it:

> To engage in a practice, to perform those actions specified by a practice, means to follow the appropriate rules. If one wants to do an action which a certain practice specifies then there is no way to do it except to follow the rules which define it. Therefore, it doesn't make sense for a person to raise the question whether or not a rule of a practice correctly applies to his case where the action he contemplates is a form of action defined by a practice. If someone were to raise such a question, he would simply show that he didn't understand the situation in which he was acting. If one wants to perform an action specified by a practice, the only legitimate question concerns the nature of the practice itself.

Once one has committed to a practice, he is bound to follow the corresponding rules because this is what being committed to a practice precisely means. To do otherwise would actually indicate the absence of such a commitment and thus a rejection of the very practice itself.[10] Therefore, taking the internal point of view with respect to a rule or a set of rules means that one commits (at least counterfactually) to the corresponding practice and – *under this assumption* – reflects on the reasons and justifications for following the rule(s), i.e. for behaving in a certain way in the context of an institutionalized practice.

In developing their rules-in-equilibrium account, Hindriks and Guala (2015) make a distinction between what they call 'observer-rules' and 'agent-rules'. This roughly corresponds to Rawls's two concepts of rules: observer-rules are formulated to summarize agents' behavior, while agent-rules by agents themselves to summarize and to guide her own behavior. However, this terminology may be regarded as slightly misleading, because it seems to indicate that an observer

[9] Hitting a homerun, scoring a three-point basket or checkmating the opponent's king are all instances of actions that are made logically possible by underlying rules defining the corresponding practices.

[10] As noted above in the case of promise-keeping, one may have an excuse, or more generally a justification, for not following some rules constitutive of a practice. But this excuse must itself be formulated in terms of rules defining the practice.

cannot take the perspective of an agent. But, in the same way that I can infer someone else's intentional attitudes without being 'in her mind', I as an external observer can also adopt the internal point of view of rules to elucidate the reasons underlying the rule-following behaviors. The crux is thus not whether one is an observer or they take part in a practice, but rather on the underlying conception of rules that one adopts. If Rawls's distinction between the justification of a practice and the justification of an action falling into a practice is considered as relevant, then the practice conception of rules seems to be the only interesting one, as it is also the only one to consider the internal point of view.

It is therefore worth noting that the BRE account is able to capture the internal point of view of rules. This can be inferred from asking what leads a set of agents to implement a given behavioral pattern, i.e. a particular equilibrium. The normativity constitutive of the internal point of view surfaces through two forms of reasoning. First, agents determine what they ought to do by forming expectations about others' behaviors and attitudes, given their broadly conceived interests. This *prudential reasoning* is essentially instrumental and is well captured by standard rationality assumptions used in a game-theoretic framework. Second, agents determine what they ought to do by reflecting on what is mandatory, permissible or forbidden. This *deontic reasoning* can be accounted for within a game-theoretic framework in a variety of ways, either or both at the level of expectations and preferences. Agents may indeed form normative expectations with regard to both theirs' and others' behaviors. Their deontic attitudes are then reflected in their preference orderings which, as a consequence, reflect more than the agents' interests in a narrow sense. This approach is especially adopted by Christina Bicchieri (2006) in her theory of social norms, but it can also be found, for instance, in Sugden's (2000b) game-theoretic account of resentment aversion, or in Rabin's (1993) theory of reciprocity.

Two remarks should be made regarding the conjunction of prudential and deontic reasoning. First, it should be noted that both forms of reasoning – and thus rule-following behavior – may call for the ability to do more or less complex inferences and correspondingly to acquire beliefs over everyone's beliefs. Strategic reasoning in games can indeed be captured by explicitly formalizing each player's beliefs about others' choices and beliefs. In particular, an epistemic interpretation of game theory takes game theory to be the extension of decision theory to strategic interactions. In this case, a game is always studied from the perspective of a player, given what she knows and believes (Perea 2012). This epistemic approach allows to state minimal conditions about the players' beliefs and reasoning for considering that they are indeed following rules. As an illustration, Lewis's (2002) theory of conventions postulates that the existence of conventions (a particular form of rules) entails that there is a common reason to believe that everyone in the relevant population follows the convention. Alternatively, Hédoin (2017) identifies institutions to 'rules-governed games' satisfying conditions of symmetric reasoning and minimal awareness. Though the details are not relevant here, what matters is that the internal point of view of rules is naturally taken once one adopts an epistemic game-theoretic framework, because it is then that beliefs and reasoning modes have to be explicitly formalized.[11]

Second, both prudential and deontic reasoning may entail reasoning over counterfactuals. The BRE account, thanks to this epistemic approach, also emphasizes the importance of the internal point of view with respect to the role played by conditional – and especially counterfactual – reasoning (Hédoin 2019). The external point of view is almost by definition blind when it comes to the importance of conditional reasoning for rule-following behavior. Indeed, while individuals following rules have to reason about counterfactuals (what would happen if one was to move their pawn backward in a chess game?), such reasoning cannot be directly revealed by behavioral patterns. In particular, reasoning about counterfactuals may give prudential reasons to act in some way. Counterfactuals may appear to be less relevant for deontic reasoning – especially if the reasoning takes the form of categorical, rather than hypothetical imperatives. But counterfactual reasoning may be needed to reflect over what Gaus (2019) calls the 'eligible moral space' – the space of morally permissible acts within a rule system. In a recent paper, Gaus and Nichols (2017) distinguish between two types of rule systems: permissive and prohibitory. The former instructs agents what they are allowed to do, while the latter only specifies what agents are prohibited from doing. Both systems allow for a great diversity of acts, though permissive systems are more restrictive than prohibitory systems. This implies that an agent contemplating to behave in a certain way, which is not explicitly prohibited or permitted by the current rules, should conditionally reason about the deontic status of the corresponding act. But even an agent intending to follow rules may have to reason over counterfactual possibilities of behaving otherwise, in order to *justify* her behavior to others.

I therefore argue that the BRE account of institutions allows to capture an important characteristic that distinguishes social kinds from natural kinds. There is more than that however: the BRE account *points to* the importance of the internal point of view to study institutions. Though other recent naturalistic social ontologies also underline a similar distinction between two perspectives or points of

[11] Though most applied game-theoretic studies of institutions do not adopt the epistemic approach, it is still possible, at least in some significant cases, to translate them into epistemic game-theoretic models. See Hédoin (2017) for an illustration using Greif 2006 study of the thirteenth century English communal responsibility system.

view, they fail to explore its implications fully. This is the issue to which I turn now. I argue in the next section that acknowledging the significance of the internal point of view has a significant implication regarding the nature of social kinds: the fact that they are characterized by a particular form of dependence with respect to individuals' intentional attitudes.

4 From Constitutive to Normative Dependence: Why Institutions Cannot be Defined by their Functions

A key issue in social ontology concerns the existence and the nature of social kinds. A contemporary definition of kinds is that of homeostatic property clusters (Boyd 1991), i.e. packages of highly correlated properties, the correlation of which is relatively stable and results from some causal mechanisms. One can think of many types of kinds. A key property of *real* kinds is that they are projectable, in the sense that they support inductive inferences and generalizations (Guala 2016). This makes them particularly suitable to scientific investigations through a variety of methods. Within real kinds, several philosophers and social scientists have argued further for a distinction between *natural* and *social* kinds on the basis of different criteria. Hacking (2000), for instance, suggests that what makes social kinds special, relatively to natural kinds, is that the former but not the latter are 'interactive' or 'reflexive'. In other words, kinds studied by social sciences are 'moving targets' that change due to the very classifications and investigations they are the object of. This peculiar relationship between social kinds and individuals' attitudes about them is sometimes referred to as a relation of 'constitutive dependence'. On this account, an entity X belongs to a kind K if and only if the members of the relevant population hold the appropriate mental states that X belongs to K. In many influential social ontologies, like Searle's (1997), the appropriate mental states are related to a form of collective acceptance which can be abbreviated in the following way:

> *Constitutive Dependence*: For all X, X is K if and only if there is a set of conditions C such that (a) there is collective acceptance CA that X is K if C, and (b) C holds.

Let call the 'dependence thesis', the view that social kinds are characterized by such a relationship between a kind and individuals' attitudes about it. Guala (2016) argues against the dependence thesis on the basis of his rules-in-equilibrium account of institutions combined with a functionalism about institutions.

Guala's (2016) rejection of the dependence thesis is grounded in the claim that each *particular* institutional practice is based on a set of rules, which can be characterized independently of people's classifications. Guala also claims that we can identify *general* forms of institutions on the basis of their *functions*. Guala argues that these functions cannot be discovered by merely participating in a social practice. Establishing the functions of an institution requires empirical knowledge that can only be obtained through scientific methods. Guala's functionalism about institutions, if it succeeded, would entail the rejection of the dependence thesis because then it would be possible to determine the nature of an institution by identifying its objective (i.e. mind-independent) functions and thus to assess the truth value of people's beliefs about them. However, in this section, I want to argue that Guala's functionalism hides the fact that there is at least another key attribute separating social kinds from natural kinds, i.e. the fact that the former are inherently *normative*. By this, I mean that social kinds can be ascribed to a particular normative status by individuals who directly take part in them. This is of great significance because, as I show below, the normativity of social kinds establishes a special dependence ('value dependence') between persons' attitudes about a particular institution and the nature of this institution, in particular in determining its essential rules. This normative or value dependence is directly related to the significance of the internal point of view, discussed in the preceding section. As we have seen, taking the internal point of view entails the combination of prudential and deontic forms of reasoning. Normativity transpires in both, and especially in the latter. The way individuals are reasoning in an institutional practice then directly reflects their normative views about the nature of the institution and the status of the rules that constitute it. On the basis of the BRE account, I sketch an approach according to which the nature of an institution is determined by which rules are actually considered as 'essential' in a meaningful sense.

Guala's defense of functionalism has several building blocks. The first is a distinction between *type-institutions* and *token-institutions*. A token-institution T_i is a particular, spatiotemporally located instantiation of a type-institution T. That a particular token T_i belongs to a type T follows from the fact that T_i implements a particular solution (i.e. equilibrium) to a more general coordination problem: "What all the (token-) institutions share is that they are solutions to the same problems, or equilibria of the same class of games" (Guala 2016: 196). A second key building-block is related to Guala's endorsement of externalism about meaning. Externalism builds on the distinction between the extension of linguistic terms and the beliefs, or other attitudes, that people have with respect to the (supposed) extension of these terms. The idea is that folk classifications of objects or events captured by linguistic terms do not have to (and generally do not) correspond to the way the world actually is. Scientific knowledge is in this perspective a better (though still approximate) way to determine the extension of linguistic terms, yet it is also fallible. The point is that the extension of linguistic terms is independent from any form of knowledge or beliefs created and held by humans.

I accept one of Guala's building blocks, externalism about meaning, as long as it is properly understood.[12] I shall, however, reject the other building block on the distinction between type- and token-institutions. A key argument in Guala's defense of functionalism is that, for any type-institution T, the historical forms taken by the various token-institutions T_1, T_2, T_3… cannot serve as a basis for characterizing what T is or should be. Guala (2016: 196) makes it clear in the case of marriage:

> The inference from practice (all marriages are heterosexual) to theory (marriage is heterosexual) trades on a confusion between types and tokens. The institution of marriage in the West, or in any historically existing culture for that matter, is not marriage in general… So by studying marriage practices in Florence during the thirteenth century, say, we can only learn about the particular way in which Florentine people organized child-rearing, reproduction, and economic cooperation at a particular time.

Arguably, insisting that token-institutions determine the content and nature of type-institutions would condemn one to some form of traditionalism forbidding the evolution of any institution (think of same-sex marriage). Moreover, on the BRE account, there is indeed a sense according to which a type-institution could be identified to a 'class of (social) games' and thus solving a kind of coordination problems. The problem with this argument is that while it may be true that the way people represent a type-institution T at a given time and location through a token-institution T_i cannot define what T is, it remains to determine how the functions of T are to be established. Another way to state the problem is the following: while one (especially the social scientist) may legitimately identify T with a class of games it solves, thus determining its functions, it is not clear why we could not identify T with another (not necessarily mutually exclusive) class of games. Functionalism supposes that we can identify the functions of an

[12] Externalism about meaning holds as long as we agree with Guala that constitutive dependence is not a characteristic of social kinds. Then, the extension of any linguistic expression designating an institution is fully determined by what individuals are doing, i.e. the equilibrium or set of equilibria in the BRE account. This is not in contradiction with the claim that what individuals are doing depends on their attitudes, including their normative attitudes about their practice.

institution, but this very identification may well be grounded in collective representations and acceptance.[13]

It is highly likely that the only way to solve this conundrum is simply to give up the type/token distinction. Fortunately, there are good arguments for doing so. In particular, as Aydinonat and Ylikoski (2018) note, it is not clear that the "thirteenth-century Florentine marriage" is a token of the type-institution "marriage". Actually, the very concept of marriage is a category issued from some scientific classification which, while it may be useful to account for social practices, is not deemed to be relevant *because* it is supposed to capture the essence of general institutions. The value of scientific concepts and theories does not lie in the fact that they approximate the nature of things they target, but rather in the fact they permit to solve well-identified problems about them. From an internal point of view, what is relevant is that the "thirteenth-century Florentine marriage" corresponds to a well-identified set of rules that individuals have *actually* followed for specific reasons. It is possible to account for these rule-following behaviors (through the BRE account for instance) and this is what makes thirteenth-century Florentine marriage more than a mere scientific concept. It could be argued that the marriage type-institution could be similarly identified by looking at the intersection of the rules that are shared by all marriage token-institutions. This minimal set of rules would then correspond to the essence of marriage as a type-institution. But there is no guarantee that the intersection is not empty, and even if this is not the case, any decision regarding what constitutes a token-marriage remains somewhat arbitrary. Ultimately, these difficulties simply highlight that the concept of marriage, like many other concepts used in the social sciences to capture institutional facts and practices (money, racism, property…), is a family-resemblance term that is not amenable to a general definition (Aydinonat and Ylikoski 2018).

While giving up the type/token distinction implies that any functionalist view of institutions is doomed, there is a way to identify *within any particular institution* what can be characterized as its 'essential' rules. This identification may then permit to characterize the nature of an institution not by its functions but by its essential rules. There are probably many ways to distinguish these two kinds of rules, but here I want to propose a criterion which is mind-dependent and objective

[13] An editor of this journal suggested that the fact that there are commonalities between people's preferences may account for the possibility of identifying type-institutions with particular functions. This is an interesting point to which I think we could reply by the following two remarks: first, these commonalities may be only contingent, meaning that may not remain stable; second, while some of these commonalities may have genetic/biological origins, others are plausibly the result of cultural mechanisms which are themselves shaped by the institutional environment (and so by specific token-institutions).

at the same time. A short example will be useful to illustrate the key ideas. Suppose an afternoon two friends meet to play chess together. As they are preparing to start the match, they discuss about the amount of time that should be allowed for each player to play their moves. Initially, they do not agree, but they quickly converge upon the duration. Before starting, one of them makes an additional suggestion to randomly place each other's pieces on their respective first rows of the chessboard.[14] However, the other vehemently disagrees and ultimately responds "this is not how we play chess" just before leaving the table.

What has happened during this short sequence? It appears that the disagreement over the amount of time did not compromise the match, in contrast to the disagreement over the initial positioning of the pieces. A plausible interpretation is that the former disagreement concerns a rule that is peripheral, or of secondary importance, for the players. They are thus ready to make concessions. The latter disagreement concerns, however, a rule that tends to be taken by most chess players as essential in their social practice. Technically speaking, with reference to the BRE account, the rule concerning the pieces' initial disposition, corresponds to an equilibrium and to beliefs that are ascribed to a particular normative significance, which the rule concerning the amount of time does not have. In other words, while both rules are related to focal points, one of these has a specific normative meaning that the other one lacks.

It is not difficult to extend this example to social institutions like money or marriage. The issue here is of course to determine where the normative dimension of some focal points comes from. It is vain to search the origins of this normative dimension in putative functions that would be exogenously ascribed by the observer to the institution. After all, these are individuals participating in the institutional practice who, taking the internal point of view, ascribe to rules a particular normative force. It is actually highly likely that no general answer to this problem is available and that whether or not a rule has a particular normative force is a matter of contingency. What can be asserted, however, is that this normative force *is* mind-dependent. This normative (or value) dependence of institutions upon individuals' attitudes is related to the two forms of justifications underlined by Rawls (1955), which I discuss in section 3. On the one hand, the normative status of a rule may result from its contribution to the overall value of the practice it is constitutive of. "Overall value" here refers to the set of reasons an individual may have for participating in an institutional practice, i.e. what makes her participation valuable according to her. There may be a great variety of reasons for engaging in an institutional practice: some of them may be intrinsic ("playing chess is

[14] This variant of chess is known as 'Fisher chess'. While this variant is officially recognized, its status with respect to "classical" chess remains disputed.

recreative and I enjoy it") and others instrumental ("playing chess helps me to improve my concentration"). In general, while each participant may have her own personal and subjective reasons for taking part in an institutional practice, we may expect that these reasons partially overlap. In this perspective, since a rule is most of the time followed as part of a hierarchically-ordered set of rules, its importance is likely to be evaluated against this whole set. What characterizes an 'essential' rule then is the fact that the whole institutional practice would lose most of its value (from the perspective of the participants themselves) were the rule not to be followed.[15] In the case of games like chess (and more specifically the practice of competitive chess), some rules may be regarded as more essential than others, because giving them up would compromise the very point of the practice as judged by the participants themselves (e.g. introducing too much randomness while the point of the practice being competition on the basis of players' computational abilities). In the case of (token) institutions of marriage, the different-sex rule may be ascribed to a particular normative significance in a society where marriage as an institution is foremost viewed as a way to lessen the risk of conflict between rival families.[16]

On the other hand, the normative status of a rule may also depend on the justification for following it *within* the practice. This will particularly be the case if the rule is regarded as promoting independent values. A rule may be regarded as particularly important because it fosters equity among the participants or helps to realize any particular moral principle. A rule may also be considered as playing a fundamental role in ensuring coordination and cooperation. While all rules and institutions presumably have the function of solving cooperation problems, some of these problems may be considered as more urgent, thus granting the corresponding rule a particular normative status.

It is interesting to note that the factors determining the normative status of rules can be couched in a functional language. Indeed, in a recent contribution Hindriks and Guala (2019) argue that institutions (and thus rules) fulfill 'etiological' and 'teleological' functions. The former helps explain why an institution

[15] Marmor (2009: 39) makes a very similar point: "engaging in the practice constituted by [the rule] *S* is valuable (at least for those who engage in it) in ways in which it *could not have been valuable without the existence of S*".

[16] Note that it does not matter whether in such a society the institution of marriage *really* contributes to regulating conflicts between families or clans. What matter is that individuals think this is the case and because of this they grant the rule a particular normative value. But since here the underlying deontic reasoning is hypothetical rather than categorical, a scientific (empirical) demonstration that marriage does not actually regulate conflicts may help people to change their normative perspectives. Obviously, scientific arguments would have less grip in the case where the deontic reasoning took the form of categorical imperatives.

continues to exist. It refers to the ability of institutions to promote cooperation within a population. The latter concerns what the institution is good for, i.e. what kind of values it promotes. Hindriks and Guala go on to argue that while the etiological function is essentially explanatory, the teleological function is mostly evaluative. Now, as far as cooperation may be regarded as a value, or more likely as being instrumentally useful in promoting other values, it may also be related to the teleological function of an institution.

I have no qualm about characterizing the normative or value dependence of institutions in terms of (teleological) function. However, in their aforementioned paper, Hindriks and Guala retain the type/token distinction that I have claimed should be rejected. Moreover, they assert that an adequate theory of institutions should not be 'moralized', i.e. institutions should not be defined in terms of the values they (are supposed to) promote.[17] If we were to accept the type/token distinction, this should probably be true for type-institutions. But the above analysis suggests that the concept of type-institution is problematic. And as my discussion of essential rules indicates, a full characterization of token-institutions actually requires the 'moralization' Hindriks and Guala are rejecting. However, this claim should not be misinterpreted. Naturalism should not be interpreted or practiced as a disguised form of *scientism*. Scientism is the view that only scientific knowledge is legitimate and that all kinds of issues and problems (including moral ones) can be solved by scientific methods. In other words, scientism implies that all philosophy can be reduced to science. This is not what naturalism means: for naturalists, science and philosophy are rather similar intellectual projects, such that the methods and theories of science can *help* to answer philosophical issues. In particular, while a naturalist may distinguish folk classifications and theories from scientific ones, she does not (or should not) forget that scientific theories and classifications are still representations, the value of which is derived from their usefulness to solve specific problems. That means that even if scientists may take as a convention that scientific concepts, which have proved useful, refer to 'real' objects (this is sometimes called 'scientific realism'), this should not be conflated with (ontological) realism. This is important for at least two reasons: first, as explained in the paper, ignoring this point runs the risk of searching in vain for 'types' and their properties. Second, a more insidious danger with this conflation is to give an inappropriate *normative significance* to scientific concepts. This is obvious with scientific concepts like 'marriage' or 'racism': while it is scientifically legitimate to define the concept of racism or marriage in a specific way to study

[17] I emphasize that the term 'moralization' has a specific meaning here. It is more often use to characterize actions as right an wrong. I am using the term as indicated in the text, following Hindriks and Guala.

specific social phenomena, these definitions cannot and should not be used as normative arguments, especially in public debates. While it may be true that there is no reason for restricting the scientific concept of marriage to heterosexual marriages, this cannot be used as an argument to claim that, in some specific society, it is ‘just’ or ‘good’ to extend marriage to same-sex persons.[18] To do otherwise would not constitute naturalism but rather authentic scientism. The same point applies to my claim that (token) institutions should be ‘moralized’. Deducing that specific values and rules have a particular normative significance *in general* from the fact that given token institutions can be characterized by a specific subset of essential rules realizing these specific values, has nothing to do with naturalism. The moralization of token institutions only entails that people’s normative views help to characterize an institution and thus should be part of any plausible *scientific* theory of institutions. But it does not (and cannot) make normative claims about which rules should regulate any institutional practice.

5 Conclusion

I have started this essay with contrasting two approaches of social ontology. According to *foundationalism*, social ontology ‘comes first’: it provides social sciences with the foundations for studying social objects and phenomena. In this perspective, social sciences must have the social ontology right before starting any theoretical or empirical investigation. *Naturalism,* quite the contrary, builds on the postulate that social sciences do not need foundations prior starting their work. Moreover, as far as there are relevant and interesting ontological issues related to the nature of the social world, naturalism holds that social sciences can potentially bring answers or at least elements helping to answer those issues. The BRE account of institutions I have presented in this paper clearly follows naturalistic lines. I have intended to show that this account could help to advance the debates over the nature of institutions. Ultimately, my main claim is that functionalism about institutions is impossible, since this would necessitate a rigid type/token distinction of institutions that no scientific account (at least not the BRE) seems able to hold.

It may be worth concluding this paper by pointing out that this claim, even if accepted, should not lead one to conclude on the impossibility of a general theory of institutions. In a scientific perspective, building classifications of type-institutions is perfectly legitimate, as long as this helps to solve well-identified problems. Hence, not only there is room for scientific theories of institutions like money, family or racism, but reflecting on the properties of institutions in general

[18] Of course, the converse is also true.

(as in the BRE account or in Guala and Hindriks's rules-in-equilibrium account) may also be scientifically valuable. However, social kinds, contrary to natural kinds, are related to values. Any theory of institutions, and the scientific image of social kinds more generally, is bounded by this value dependence. A naturalist (rather than a 'scientist') social ontology should give scope, in its endeavor to characterize the nature of particular institutions, to the role played by values endorsed by persons and that ground the particular normative importance given to some rules.

Acknowledgments: This paper has benefited from the comments and remarks of several anonymous referees. I would like to thank them, as well as an editor of this journal, as they have greatly contributed to improve the quality of this work. I also thank the participants in seminars at the University of Paris 1 Sorbonne, at the Ecole Normale Supérieure of Lyon, and at the University of Reims Champagne-Ardenne, where different versions of this paper have been presented. Finally, I thank Eva Barteková for her remarks and careful reading of the final version. All errors and omissions remain mine.

References

Aydinonat, N. E., and P. Ylikoski. 2018. "Three Conceptions of a Theory of Institutions." *Philosophy of the Social Sciences* 48 (6): 550–68.

Basu, K. 2018. *The Republic of Beliefs: A New Approach to Law and Economics*. Princeton, NJ: Princeton University Press.

Bicchieri, C. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

Binmore, K. G. 1998. *Just Playing: Game Theory and the Social Contract*. MIT Press.

Boyd, R. 1991. "Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 61 (1/2): 127–48.

Epstein, B. 2015. *The Ant Trap: Rebuilding the Foundations of the Social Sciences*. Incorporated: Oxford University Press.

Gaus, G. 2019. "Morality as a Complex Adaptive System: Rethinking Haye's Social Ethics." In *The Oxford Handbook of Ethics and Economics*, edited by M. D. White Oxford: OUP.

Gaus, G., and S. Nichols. 2017. "Moral Learning in the Open Society: The Theory and Practice of Natural Liberty." *Social Philosophy and Policy* 34 (1): 79–101.

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.

Guala, F. 2016. *Understanding Institutions: The Science and Philosophy of Living Together*. Princeton, NJ: Princeton University Press.

Hacking, I. 2000. *The Social Construction of What?*, Revised edition. Cambridge, Mass: Harvard University Press.

Hart, H. L. A. 2012. “Leslie Green.” In *The Concept of Law. Clarendon Law Series*, 3rd ed., edited by R. Joseph, and P. A. Bulloch. Oxford, New York: Oxford University Press.

Hédoin, C. 2014. “A Framework for Community-Based Salience: Common Knowledge, Common Understanding and Community-Membership.” *Economics and Philosophy* 30 (03): 365–95.

Hédoin, C. 2017. “Institutions, Rule-Following and Game Theory.” *Economics and Philosophy* 33 (1): 43–72.

Hédoin, C. 2019. “Institutions, Rule-Following and Conditional Reasoning.” *Journal of Institutional Economics* 15 (1): 1–25.

Hindriks, F., and F. Guala. 2015. “Institutions, Rules, and Equilibria: A Unified Theory.” *Journal of Institutional Economics* 11 (03): 459–80.

Hindriks, F., and F. Guala. 2019. “The Functions of Institutions: Etiology and Teleology.” *Synthese*. https://doi.org/10.1007/s11229-019-02188-8.

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

Hurwicz, L. 2008. “But Who Will Guard the Guardians?.” *American Economic Review* 98 (3): 577–85.

Lewis, D. K. 2002. *Convention: A Philosophical Study*. Malden, MA: John Wiley and Sons.

Marmor, A. 2009. *Social Conventions: From Language to Law*. Princeton, NJ: Princeton University Press.

Myerson, R. B. 2009. “Fundamental Theory of Institutions: A Lecture in Honor of Leo Hurwicz.” *Review of Economic Design* 13 (1): 59.

Perea, A. 2012. *Epistemic Game Theory: Reasoning and Choice*. New York: Cambridge University Press.

Rabin, M. 1993. “Incorporating Fairness into Game Theory and Economics.” *The American Economic Review* 83 (5): 1281–302.

Rawls, J. 1955. “Two Concepts of Rules.” *The Philosophical Review* 64 (1): 3.

Schelling, T. C. 1981. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Searle, J. R. 1997. *The Construction of Social Reality*. New York, NY: Simon and Schuster.

Searle, J. R. 2010. *Making the Social World: The Structure of Human Civilization*. Cambridge, MA: Oxford University Press.

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 (01): 1–22.

Sugden, R. 2000a. “Team Preferences.” *Economics and Philosophy* 16 (02): 175–204.

Sugden, R. 2000b. “The Motivating Power of Expectations.” In *Rationality, Rules and Structures*, edited by J. Nida-Rümelin, and W. Spohn, 103–29. Dordrecht, The Netherlands: Springer.

Sugden, R. 2005. *The Economics of Rights, Cooperation and Welfare*, 2nd ed. New York, NY: Palgrave Macmillan.

Sugden, R. 2011. “Salience, Inductive Reasoning and the Emergence of Conventions.” *Journal of Economic Behavior & Organization* 79 (1–2): 35–47.

Sugden, R. 2016. “Ontology, Methodological Individualism, and the Foundations of the Social Sciences.” *Journal of Economic Literature* 54 (4): 1377–89.

Tuomela, R. 2013. *Social Ontology: Collective Intentionality and Group Agents*. Oxford, New York: Oxford University Press.

Wittgenstein, L. 1965. *The Blue and Brown Books*. New York, NY: HarperCollins.