Journal of Institutional Economics (2019), 15: 1, 1–25
 doi:10.1017/S1744137418000073
First published online 13 March 2018

Institutions, rule-following and conditional reasoning

CYRIL HÉDOIN*
Economics and Management Research Center REGARDS, University of Reims Champagne-Ardenne, Reims, France

Abstract. This paper is a contribution to the advancement of a naturalistic social ontology. Individuals participate in an institutionalized practice by following rules. In this perspective, I show that the nature, the stability, and the dynamics of any institution depend on how people reason about states of affairs that do not occur. That means that counterfactual reasoning is essential in the working of institutions. I present arguments for the importance of counterfactuals as well as a game-theoretic framework to account for them. Since the role of counterfactuals does not directly transpire in people's behavior, the whole discussion can be seen as a broad argument against behaviorism in philosophy and the social sciences.

1. Introduction

Social ontology can be defined as the inquiry into the nature of social reality and especially of institutions, i.e. sets of norms, rules, conventions, and organizations that regulate established practices in a given population. Recent contributions by economists and philosophers have attempted to provide naturalistic foundations for a theory of the nature of institutions. They share a willingness to develop an ontological account of institutions through the tools, theories, and methods of the social sciences, more specifically of economics.

My aim in this paper is to expand this general endeavor to give naturalistic foundations to social ontology, especially on the basis of economics and game theory. I will do so by focusing more particularly on the role played by agents' *conditional reasoning* in the functioning of institutions. Conditional reasoning typically takes the "if . . . , then . . . " form where a (set of) antecedent(s) is related to a (set of) consequent(s). I propose to investigate the kinds of conditional reasoning that agents must use when participating in an institutional practice. I assume that participation in an institutional practice occurs through *rule-following behavior*. In other words, to participate in an institutional practice consists in following a rule or a set of rules. In turn, an agent follows a rule in performing an action only if he acts on the basis of the appropriate intentional attitudes (including the intention to follow the rule) formed on the basis of some shared inferences. I shall argue that the reasoning process of individuals

*Email: cyril.hedoin@univ-reims.fr

participating in an institutionalized practice and leading to these inferences must rely on both *indicative* and *subjunctive* conditionals. More specifically, I intend to show that the nature, the stability, and the dynamics of any institution depend on how people reason about states of affairs that *do (did, will) not occur*. That means that counterfactual reasoning is essential in the working of institutions. I present arguments for the importance of counterfactuals as well as a game-theoretic framework to account for them. The latter uses the resources of epistemic game theory (Brandenburger, 2014; Perea, 2012) and of semantic models of epistemic logic (Bacharach, 1993; Stalnaker, 1996).

The rest of the paper is organized as follows. Section 2 lays the main motivations of the paper and briefly discusses the relevance of naturalistic accounts of institutions. In section 3, I present a rule-following account of institutions similar to Hindriks and Guala's (2015) that I augment with a formal game-theoretic framework. I propose to characterize an institution as a set of *game situations*, which are themselves formalized as semantic epistemic models of games. Section 4 proposes a brief overview of the philosophical literature on conditionals and distinguishes different kinds of counterfactuals (causal and epistemic) that must be taken into account by the agents' reasoning process. Sections 5 to 7 offer three arguments for the importance of subjunctive conditionals and more specifically counterfactuals in order to understand the nature of institutions: (1) counterfactuals are needed to characterize both equilibrium behavior and the agents' rationality; (2) counterfactuals directly determine off-the-equilibrium-path behavior and thus determine the dynamics of institutions; (3) the capacity to adopt counterfactual reasoning is essential for individuals' ability to simulate others' reasoning in the context of coordination problems. Section 8 briefly concludes. An online appendix provides the formal details of the framework developed.[1]

2. The relevance of naturalistic accounts of institutions

Recently, several economists and philosophers have developed naturalistic accounts of the nature of institutions and institutional facts. More specifically, J. P. Smit, Filip Buekens, and Stan du Plessis have developed an "incentivized-action view" of social reality in several writings (Smit *et al.*, 2011, 2014, 2016). Smit *et al.* argue that social reality and especially what they call "institutional objects" can be accounted for in terms of actions and their underlying incentives. Philosophers Francesco Guala and Frank Hindriks have also developed a naturalistic account of the nature of institutions (Guala, 2016; Guala and Hindriks, 2015; Hindriks and Guala, 2015). Distinguishing

1 The appendix is available at this address: https://sites.google.com/site/cyrilhedoin/Institutions%2C%20Rule-Following%20and%20Conditional%20Reasoning.pdf?attredirects=0&d=1 (accessed January 26, 2018).

https://doi.org/10.1017/S1744137418000073 Published online by Cambridge University Press

between an "institutions-as-rules" and an "institutions-as-equilibria" approach in philosophy and the social sciences (especially economics), they argue for a broadly synthetic view that they call "rules-in-equilibrium."

Smit *et al.*'s and Guala and Hindriks's accounts share many features. The most significant feature is that they both endorse a form of *naturalism* in social ontology. By naturalism, I mean the view that issues in ontology (social or whatever) can only be solved through a systematic reliance on scientific methods, theories, and results. The second significant feature shared by the two accounts is the fact that they both build largely on the tools and concepts provided by economics and game theory. Smit *et al.*'s incentivized-action view is nothing but a straightforward extension of microeconomic reasoning to social ontology. Institutions are identified as behavioral patterns resulting from incentives that can be related to given institutional objects (e.g. patterns generated by traffic lights on roads). On the other hand, Guala and Hindriks's rules-in-equilibrium account explicitly builds on a game-theoretic framework and more specifically on Herbert Gintis's (2009) account of social norms as correlated equilibria. The significance of this naturalistic endeavor lies in the fact that social ontology until now has consisted mostly of conceptual analysis with little if any interaction with the social sciences (Guala, 2007). A naturalistic social ontology builds on the principle that ontological investigation is actually no different from science. In particular, its ultimate aim is to permit the production of theoretical knowledge that can be confronted with empirical evidence. The use of economic theory and game theory is especially relevant in this perspective.

The fact of using the scientific methods and tools of a particular discipline is not without risk. The way economists have studied institutions, especially through a game-theoretic framework, has sometimes reflected a form of behaviorism that continues to inspire modern economics.[2] Behaviorism has at least two implications, one methodological, the other ontological. On the methodological side, it leads to a restriction of the relevant informational basis for theory building and theory testing to choice-data. Other kinds of information (obtained for instance through introspection, interviews, or brain scans) are regarded as irrelevant. On the ontological side, it encourages the adoption of reductionist or even eliminativist views of social objects such as institutions. Though the inference from the claim "only choice-data are relevant" that "there is nothing beyond choice behavior" is not logical, it is nonetheless regarded as highly natural by many social scientists (List and Spiekermann, 2013). This point is particularly relevant for institutional analyses based on a game-theoretic framework because it contributes to explaining the prevalence of the "institutions-as-equilibria" view of institutions among economists. As a game-

2 See especially the methodological manifesto of Gul and Pesendorfer (2008). Reflections over the form of behaviorism that underlies contemporary mainstream economics, especially revealed-preference theory, can be found in Dietrich and List (2016), Hands (2013), and Ross (2014).

https://doi.org/10.1017/S1744137418000073 Published online by Cambridge University Press

theoretic equilibrium is defined as a strategy profile with specific properties, this encourages us to regard institutions as being mere behavioral patterns and to downplay the importance of underlying factors responsible for the observed behavior.

Though it cannot be straightforwardly characterized as behaviorist, Smit *et al.*'s incentivized-action view identifies institutions with behavioral patterns and the underlying incentives responsible for them. Therefore, this account does not regard the forms of reasoning that underlie institutions as being constitutive of them. Basically, it consists in claiming that any behavioral regularity *R* is an institution as soon as we have been able to identify it with a set of incentivized actions. There are two problems with this view. First, as there are behavioral regularities in every animal population and since most if not all organisms behave in a more or less consistent manner, institutions are not specific to humans. Second, while it may be perfectly right to claim that all animal species have institutions if we take an *observer perspective*, this is not satisfactory from an *agent perspective*, i.e. if our aim is to explain behavioral patterns through the agents' reasoning and motivations. Here, we have strong empirical reasons to consider that humans are highly specific, starting with their ability to develop devices to store knowledge and communicate (language, tools) and their cognitive ability to use abstract reasoning and to form complex and nested intentional attitudes (Tomasello, 2014).

This last point is recognized by Guala and Hindriks in their writings. They emphasize that the working of institutions depends on how individuals simulate others' reasoning process.[3] They propose moreover a characterization of institutions in terms of a list of conditionals corresponding to a set of "regulative" rules that the agents follow. They remain however silent regarding the nature of the "if . . . , then . . . " conditionals. My main claim in this paper is precisely that a full understanding of the nature of institutions depends on this last point, and leads us to move away further from behaviorism. In particular, while Hindriks and Guala's account seems to restrict its attention to indicative conditionals, I shall argue for the importance of *counterfactual* reasoning on the basis of three complementary reasons that are developed in sections 5, 6 and 7. First, the way in which individuals reason about counterfactuals matters for the nature of equilibrium play. Second, the way in which individuals reason about counterfactuals determines the range of parameter values for which a given strategy profile corresponds to an equilibrium in a game. Finally, recent research within the so-called field of "theory of mind" indicates that the ability to reason about counterfactuals is essential in the simulation process that individuals use to coordinate.

3 See especially Guala (2016: chapter 7). Note moreover that Guala and Hindriks also recognize the distinction between the observer-perspective and the agent-perspective that I make in the text.

https://doi.org/10.1017/S1744137418000073 Published online by Cambridge University Press

3. Institutions as rule-governed games: a formal framework

This section provides a formal framework for a rule-following account of institutions.[4] The rule-following account finds its roots in the late writings of Ludwig Wittgenstein, especially *Philosophical Investigations* (Wittgenstein, 2010). I take institutions to be sets of norms, legal rules, conventions, and organizations that generate behavioral patterns. Any institution is coextensive with a practice, so I will routinely speak in the paper of *institutional practices* to designate both behavioral patterns and the underlying rules that individuals are following. To participate in an institutional practice consists in following a (set of) rule(s). For instance, playing baseball consists in participating in an institutional practice and thus to behave according to a set of rules that are coextensive with the game of baseball. Similarly, to sell a good on a market is to take part in an institutional practice that is constitutively defined by the rules regulating the exchange. The notion of rule-following in Wittgenstein's writings is notoriously controversial but I will momentarily set aside this issue here. The formal framework I propose here nevertheless characterizes rule-following in a precise way in terms of the notion of *rule-governed games*.[5]

We start with the standard characterization of a game G as a tuple $< N, \{S_i, u_i\}_{i \in N} >$. N is the (finite) set of $n \geq 2$ players and S_i the (finite) set of players $i = (1, \ldots, n)$'s pure strategies s_i. The set of strategy profiles is denoted $S = \Pi_i S_i$. Each player is endowed with a complete and consistent preference ordering over the set of strategy profiles with u_i: $S \rightarrow \Re$ a von Neumann–Morgenstern utility function cardinally representing this ordering. Throughout the paper, I assume that the players are Bayesian rational, i.e. they choose the strategy that maximizes the expectation of their utility function given some probability measure of others' strategy choice. As it is characterized, a game provides only a partial description of a strategic interaction. To characterize an institution as a rule-governed practice, we must be able to formalize each peculiar instance or situation where a rule is followed by the members of the relevant population. We then have to be able to specify what the players' full and partial beliefs about a set of events are, what the kinds of inferences they are making on this basis, and finally what they are actually doing. Following Aumann and Dreze (2008), I will use the term "game situation" to capture the formalization of specific states of affairs where players are following a rule as part of an institutional practice.

I formalize game situations on the basis of semantic models of epistemic/doxastic logic. An epistemic/doxastic logic is a logic of knowledge and/or beliefs. It provides a language to formulate propositions and theorems about what is known and believed given a set of axioms defining knowledge

4 See Hédoin (2017) for a similar, though not identical, framework.

5 Sillari (2013) provides an interesting and helpful discussion of Wittgenstein's account of rule-following that is largely in line with the framework developed here.

https://doi.org/10.1017/S1744137418000073 Published online by Cambridge University Press

and belief operators. An epistemic/doxastic model then provides the semantic counterpart for the logic by assigning a truth value to any proposition contained in the language. I will use the term of "semantic epistemic model" (*s.e.m.* for short) to refer to the tuple M: $< \Omega, \boldsymbol{\omega}, \{B_i, C_i, p_i\}_{i \in N} >$. Ω refers to the set of possible worlds (or states of the world) x.[6] A possible world is a complete description of all the features that are deemed as relevant by the modeler and/or the decision makers. Formally, it corresponds to a list of true propositions as expressed in the underlying logic. $\Psi = 2^{\Omega}$ is the set of all subsets of Ω. Any member of Ψ is called an "event." An event is the semantic counterpart of proposition ϕ, i.e. the set of states in which ϕ is true. For any such proposition, I denote $[\phi]$ as the corresponding event. $\boldsymbol{\omega}$ denotes the actual world, i.e. what the players are actually doing, knowing and believing. In other words, $\boldsymbol{\omega}$ provides a full description of a particular way a game is played and a rule is followed. Each B_i corresponds to a binary accessibility relation indicating what the possible worlds that are consistent with what i knows/believes in any world x are. As usual, xB_iy is read as "world y is accessible from world x," i.e. y is consistent with what i knows/believes in x. I denote $\boldsymbol{B_i}$: $\Omega \rightarrow \Psi$ and $B_i(x) = \{y \in \Omega \mid xB_iy\}$ the corresponding possibility operators and possibility sets. It should be noted that B_i, $\boldsymbol{B_i}$ and $B_i(.)$ are three equivalent ways to characterize a player's epistemic/doxastic state, each indicating for any given world which subset of possible worlds a player believes to contain the actual world. Each C_i: $\Omega \rightarrow S_i$ denotes player i's decision function. It specifies what each player is doing in each possible world. I write $C(x) = (C_1(x), \ldots, C_n(x)) \in S$ to denote the strategy profile implemented in x. Finally, $\{p_i\}_{i \in N}$ is a vector of probability measures defined over Ψ. Thus each p_i defines player i's prior beliefs about events (propositions). Correspondingly, $p_{i,x}$ specifies i's full and partial beliefs in possible world x. If the players are Bayesian rational, then i's degree of belief in proposition ϕ at x is defined as:

$$p_{i,x}([\phi]) = \frac{p_i([\phi] \cap B_i(x))}{p_i(B_i(x))} \tag{1}$$

I call an *ex interim* epistemic game $\boldsymbol{G_\omega}$ any pair $< G, M >$. When the *s.e.m.* does not define the actual world $\boldsymbol{\omega}$, I call the pair $< G, M >$ an *ex ante* epistemic game $\boldsymbol{G}$, or epistemic game for short (the labels *ex ante* and *ex interim* will become clearer below).

The kinds of attitudes captured by the accessibility relations B_i depend on the properties of the latter. Following Gintis (2009), Hindriks and Guala (2015) characterize institutions as correlated equilibria, i.e. probabilistic distributions of strategy profiles such that each player maximizes her expected utility conditional

6 As it is not essential to my point, I assume for simplicity that Ω is finite.

https://doi.org/10.1017/S1744137418000073 Published online by Cambridge University Press

on the strategy she is playing. From this point of view, Robert Aumann (1987) has demonstrated an important theorem:

> **Aumann's Theorem –** For any correlated equilibrium in an arbitrary game G, there is an epistemic game $\boldsymbol{G}$ whose *s.e.m.* satisfies the two following conditions: (1) all players maximize their expected utility in each possible world $x \in \Omega$ (Bayesian Rationality); (2) the probability measures are identical, i.e. $p_1 = \ldots = p_n = \boldsymbol{p}$ (Common Prior). Conversely, for any *s.e.m.* satisfying Bayesian Rationality and Common Prior, there is a corresponding game G where the players implement a correlated equilibrium.

Aumann's theorem is interesting from the perspective of social ontology, especially given Guala and Hindriks's account. It is formulated, however, on the basis of knowledge-belief semantic structures that assume that each player is endowed with an information partition over Ω. Such structures have notoriously been deemed problematic in their treatment of knowledge, especially as they rely on the assumption that the players have introspective access to their knowledge. A more satisfactory approach consists in attributing to players only doxastic states (beliefs), which is done by allowing the possibility that individuals may be wrong. More formally, I will assume that the players' possibility sets satisfy the two following conditions:

(a) $\forall x \in \Omega$: $\exists y$: $y \in B_i(x)$ (Coherence)
(b) $\forall x, y \in \Omega$: if $y \in B_i(x)$, then $B_i(x) = B_i(y)$ (Introspection)

Condition (a) guarantees that a player's beliefs are consistent. Condition (b) indicates that each player has introspective access to her beliefs. The latter assumption is far less contentious in the case of belief than in the case of knowledge (Stalnaker, 1996). However, there is no guarantee that one's beliefs are correct, i.e. contrary to knowledge-belief structures, we do not require that $x \in B_i(x)$ for all $x \in \Omega$. As a consequence, the accessibility relations B_i do not define partitions as in Aumann's theorem. We can now define precisely the event that i fully believes $[\phi]$ in world x through a semantic belief operator B_i:

$$\mathrm{B_i}[\phi] = \{x \in \Omega \mid B_i\ (x)\ \subseteq\ [\phi]\}. \tag{2}$$

Player i's partial beliefs are similarly defined, in conjunction with the probability measure $p_{i,x}$. The event that i believes event $[\phi]$ with degree of at least p at x is:

$$\mathrm{B_i^p}[\phi] = \{x \in\ \Omega \mid B_i(x)\ \cap\ [\phi]\ \neq\ \emptyset \text{ and } p_{i,x}([\phi])\ \geq\ p\}. \tag{3}$$

Finally, define the common accessibility relation B^* as the transitive closure of the accessibility relations $\{B_i\}$, i.e. $xB^*\ y$ if and only if, denoting $x = \omega_1$ and $y = \omega_n$, for any $1 \leq m \leq n - 1$ there is a B_j such that $\omega_m B_j \omega_m + 1$. Then, the common belief operator $\mathrm{B^*}$ denotes the event that "everyone believes that

https://doi.org/10.1017/S1744137418000073 Published online by Cambridge University Press

everyone believes that ...” and so on *ad infinitum*:

$$\mathrm{B}^*[\phi] = \{x \in \Omega \mid B^*(x) \subseteq [\phi]\}. \tag{4}$$

Denote $[\mathrm{r_{i,x}}]$ the event that player i is rational in x. It is defined straightforwardly as:

$$\left[\mathrm{r_{i,x}}\right] = \{x \in \Omega \mid \mathrm{E}u_{i,x}(\mathrm{C}_i(x)) \geq \mathrm{E}u_{i,x}(s_i') \text{ for all } s_i' \in S_i\}, \tag{5}$$

where $\mathrm{E}u_{i,x}$ is i's expected utility in x by playing s_i, given her beliefs defined by B_i and $p_{i,x}$.

Denote $[\mathrm{r_x}] = \cap_{i \in N}[\mathrm{r_{i,x}}]$. It follows that the event that it is common belief that everyone is rational in x corresponds to:

$$\mathrm{B}^*[\mathrm{r_x}] = \{x \in \Omega \mid B^*(x) \subseteq [\mathrm{r_x}]\}. \tag{6}$$

On this basis, the fact that a set of players N is following a rule in some game situation is formalized by the *ex interim* epistemic game $\boldsymbol{G_\omega}$ with the following properties:

Rule-following in a game situation – The members of a population are following a rule in a game situation if there is an *ex interim* epistemic game $\boldsymbol{G_\omega}$ in which there is an event $\mathrm{R} \subseteq \Psi$ such that:

(a) For all $i \in N$: $\mathrm{R} \subseteq [\mathrm{r_{i,x}}]$ (Rationality).
(b) For all $x \in \mathrm{R}$: $\mathrm{B}^*[\mathrm{r_x}]$ (Common Belief in Rationality).
(c) For all $i \in N$: $\mathrm{B_i R}$ (Mutual Accessibility).
(d) For all $i \in N$ and all $x \in \mathrm{R}$: $p_{i,x} = \boldsymbol{p_x}$ (Agreement).
(e) $\boldsymbol{\omega} \in \mathrm{R}$ (Actuality).

According to this definition, we can say that the players are following a rule if, in the actual world, there is a common belief in everyone's rationality and the players agree on the probabilistic distribution of events, including everyone's strategy choices. Moreover, we require that each player believes that a rule is followed in this sense (Mutual Accessibility condition). It is pretty clear that the combination of Rationality, Agreement, and Common Belief in Rationality entails (with the implicit assumption that the game G is also commonly believed) equilibrium behavior by the players in $\boldsymbol{\omega}$.[7] The Mutual Accessibility condition guarantees that this equilibrium behavior is not sheer coincidence as every probability measure $p_{i,x}$ is defined for the same subset of events.

Now, for a given game G, consider a class of *s.e.m.*, differing only regarding the identity of the actual world $\boldsymbol{\omega}$. Denote as M_k each specific model of this class and R_k the corresponding event that a rule is followed in game situation k.

7 In $\boldsymbol{\omega}$, each player i's partial beliefs are defined by the probability measure $p_{i,\omega}$ on $B_i(\boldsymbol{\omega})$. Given that the decision functions C_i define each player's strategy choice at $\boldsymbol{\omega}$, each i has partial beliefs in others' strategy choices. The common belief in rationality assumption entails that everyone believes that everyone plays her best response. Rationality guarantees that no one actually wants to deviate.

https://doi.org/10.1017/S1744137418000073 Published online by Cambridge University Press

Construct then an *ex ante* epistemic game $\boldsymbol{G} = (G, M)$ where the state space Ω in M is partitioned by all the R_k.[8]

Proposition 1 – The model M of G satisfies the following conditions:

(a) For all $i \in N$: $[r_{i,x}] = \Omega$ (Rationality).
(b) For all $x \in \Omega$: $B^*[r_x]$ (Common Belief in Rationality).
(c) For all $i \in N$: $p_i = \boldsymbol{p}$ (Agreement).

To fully recover Aumann's result, say that two possible worlds $x \sim_i y$ are subjectively indistinguishable for i whenever for any possible world z we have $z \in B_i(x)$ and $z \in B_i(y)$. The binary relations $\sim_i$ then define equivalence classes that partition the state space. Because of Mutual accessibility, in $\boldsymbol{G}$ we will have $x \sim_i y$ only if x and y belong to the same event R_k. The partition of Ω according to the R_k is thus equivalent to a knowledge partition of the kind assumed by Aumann. Following Proposition 1, this gives:

Proposition 2 – In any *ex ante* epistemic game $\boldsymbol{G}$ whose state space is the union of events R_k that a rule is followed, Aumann's theorem entails that the common prior $\boldsymbol{p}$ implements a correlated equilibrium in the corresponding game G.

For the rest of the paper, I refer to any such epistemic game $\boldsymbol{G}$ as a *rule-governed game*. My aim is now to inquire into the nature and the role played by conditional reasoning in such rule-governed games.

4. Conditional reasoning and kinds of conditionals

According to the above framework, an institution or institutional practice corresponds to a set of game situations where some rule is followed. Note that I have implicitly assumed that the institutional practice consists in following a single rule in different circumstances (the various game situations distinguished by the actual world $\boldsymbol{\omega}$). Section A.1 of the appendix shows how the framework can be extended to the case of institutions with multiple rules.

The distinction between the epistemic game $\boldsymbol{G}$ and the game situations $\{\boldsymbol{G_\omega}\}$ is particularly useful to account for the difference between the observer perspective and the agent perspective of rules. The latter is particularly emphasized by Guala (2016: 54): "an observer formulates an [observer-rule] to represent or summarize others' behavior; an agent formulates an [agent-rule] to represent and to guide her own behavior." From an observer perspective, we want to be able to describe, explain, and possibly predict everyone else's behavior (or everyone's behavior if the observer is "outside" the relevant population). From an agent perspective, we want to be able to tell how an agent should reason in a given situation. As the epistemic game $\boldsymbol{G}$ provides a statistical summary of

8 Formally, this means that $\Omega = \cup_k R_k$ and $R_k \cap R_{k'} = \varnothing$ for any other game situation k'. The simplest case is where each R_k is a singleton, i.e. $\boldsymbol{\omega} = R_k$ for all k.

https://doi.org/10.1017/S1744137418000073 Published online by Cambridge University Press

Figure 1. The "Hawk–Dove" game

		Column	
		Hawk	Dove
Row	Hawk	5 ; 5	1 ; 6
	Dove	6 ; 1	0 ; 0

the behavioral pattern corresponding to the institutions (through the $\{p_i\}$), it is clearly relevant from an observer perspective. The game situations $\{\boldsymbol{G}_\omega\}$ are also useful in this perspective as they provide more specific details about each player's behavior in given circumstances. However, from an agent perspective, only the game situations seem to be relevant.

The corollary that game situations $\{\boldsymbol{G}_\omega\}$ rather than the epistemic game $\boldsymbol{G}$ are the right "unit of analysis" from the agent perspective is significant for Guala and Hindriks's account. Indeed, the latter is grounded on Gintis's (2009) claim that social norms are formally related to the common prior $\boldsymbol{p}$ in $\boldsymbol{G}$. But from an agent perspective, only the posteriors $\{p_{i,\omega}\}$ in each specific game situation $\boldsymbol{G}_\omega$ are significant. The latter, and not the former, capture the players' reasoning in an institutional practice. As I intend to show in the rest of the paper, this contributes to explaining why the kinds of conditionals used by the players in their reasoning matter in order to understand institutions.

Hindriks and Guala (2015) claim that institutions can be characterized as a set of "if . . . , then . . . " statements. These statements correspond to conditional strategies in an augmented version of game *G*. For instance, consider the following simple "Hawk–Dove" game (Figure 1).

This game is well-known among biologists and economists as it has been extensively used to explain territorial conflicts among animals and property norms among humans. Regarding the latter, a proto-institution of property corresponds to the pair of conditionals

"If incumbent, then play Hawk"
"If challenger, then play Dove"

This pair of conditionals constitutes both a Nash equilibrium in the augmented game figuring the conditional strategy (see Figure 2[9]) and a correlated equilibrium in the game in Figure 1, with an underlying *s.e.m.* satisfying all conditions of Aumann's theorem. What is the status of these conditionals and

9 Payoffs are computed by assuming that each player has a probability of ½ to be incumbent and of ½ to be challenger and that there is always exactly one challenger and one incumbent.

https://doi.org/10.1017/S1744137418000073 Published online by Cambridge University Press

Figure 2. The "Hawk–Dove" game with a conditional strategy

		Column		
		Hawk	Dove	H if I, D if C
Row	Hawk	5 ; 5	1 ; 6	3 ; 11/2
	Dove	6 ; 1	0 ; 0	3 ; 1/2
	H if I, D if C	11/2 ; 3	½ ; 3	7/2 ; 7/2

do they exhaust the content of an institution? I shall claim that this depends on the perspective (whether of observer or agent) taken.

Conditionals and their various kinds have been the object of deep and difficult debates among philosophers and logicians. I cannot pretend to provide an exhaustive summary of them here and do full justice to their many subtleties.[10] Among the three generic kinds of conditionals generally distinguished (material, indicative, and subjunctive), I will argue for the role played by subjunctive conditionals and more specifically by *counterfactuals* in the agents' reasoning. Indicative and subjunctive conditionals are properly expressed by the "if . . . , then . . . " construction. They differ in the "mood" in which they are expressed. Here is a classic example used by David Lewis (1973: 3) among others:

If Oswald did not kill Kennedy, then someone else did.
If Oswald had not killed Kennedy, then someone else would have.

That the mood in which these otherwise similar statements are expressed matters can be seen from the fact that while the first, indicative one appears to be reasonable, the second, subjunctive one is arguably doubtful. This example shows that these two kinds of conditionals differ in nature, as one statement expressed in a given mood may be true, and the same statement expressed in the other mood may be false. When their antecedent is taken to be false as a matter of fact, subjunctive conditionals are sometimes called *counterfactuals*. There is however no consensus on how to distinguish more general subjunctive conditionals from authentic counterfactuals on this basis (see Bennett, 2003: 11–12). I will thus use the two terms (subjunctive conditionals and counterfactuals) interchangeably, bearing in mind that the whole discussion applies to both of them.

Counterfactuals are generally singled out on the basis of at least two properties. The first property is that counterfactuals are *variably strict* (contrary

10 The interested reader may however consult Bennett (2003).

https://doi.org/10.1017/S1744137418000073 Published online by Cambridge University Press

to material conditionals). That means that the truth value of counterfactuals is relative to a measure of similarity between possible worlds: a given counterfactual may be true according to one measure, but false according to another one. Stalnaker (1968) and Lewis (1973) provide the standard account of counterfactuals in terms of closeness conditions allowing such a measure of similarity between possible worlds.[11]

The second characteristic of subjunctive conditionals that sets them apart from indicative conditionals concerns the kind of underlying reasoning mode used by their users. It is widely accepted that indicative conditionals correspond to evidential suppositions made under matter-of-fact reasoning, i.e. reasoning used in cases where the antecedent is *known* to be true. The natural way to deal with evidential suppositions and thus to form rational beliefs for indicative conditionals is through standard Bayesian conditioning. It follows that indicative conditionals are straightforwardly captured in the semantic framework of section 2 through the use of the probability measures $\{p_i\}$ and Bayesian conditioning denoted by expression (1). However, matters seem to be different for subjunctive conditionals. In this case, the antecedent is *not* known to be true and may indeed be known to be *false*. Counterfactual reasoning seems to be related (though perhaps not identical) with *interventional* suppositions (Bradley, 2016: 139–40). That is, counterfactual reasoning is grounded on one's full beliefs (or knowledge) about the world's causal structure. This results in a significant difference in the respective nature of evidential and counterfactual reasoning: under the former, you make the supposition holding constant your other beliefs as most as possible. Under the latter, quite on the contrary, you have to adjust your other beliefs to maintain your entrenched belief in the causal structure of the world. That counterfactuals are variably strict conditionals precisely reflects this dependence of their truth value upon some assumptions about the causal structure of the world.

Acknowledging that evidential suppositions and counterfactual reasoning sustain a distinctive kind of conditionals, we have to question what their role in the functioning of institutions is. In this perspective, it is useful to distinguish between two types of counterfactuals that agents following rules may have to take into account in their reasoning: *causal* counterfactuals and *epistemic* counterfactuals (Stalnaker, 1996: 139). The former refer to cases where an agent has to consider what the consequences would be if she were to do something she is not actually doing or going to do. This is a reasoning about what I shall call *counterfactual outcomes* or *consequences*. The latter are about the way a player would revise her beliefs were she to learn some surprising information that renders them mistaken. This is a reasoning about what I shall call *counterfactual beliefs* that is usually more related to what is generally known

11 Section A.2 of the appendix provides a formal account of Lewis's systems of spheres as a way to capture counterfactuals as variably strict conditionals.

https://doi.org/10.1017/S1744137418000073 Published online by Cambridge University Press

Figure 3. The "Prisoner's Dilemma"

		Column	
		C	D
Row	C	5 ; 5	1 ; 6
	D	6 ; 1	2 ; 2

as the problem of belief revision. In the next sections I intend to show that both types of counterfactuals are relevant to understanding the nature of institutions. In the process I will show how to incorporate counterfactuals into the semantic framework used above to characterize institutions and rule-following behavior.

5. Equilibrium behavior and the nature of rationality

In economics and game theory, rationality refers to a set of axioms related to the agent's behavior. More specifically, a rational agent is someone whose behavior conforms to the axioms of Bayesian decision theory. Broadly, a Bayesian rational agent is someone whose preferences and beliefs have properties that make them amenable to being represented by a utility function and a probability function satisfying some uniqueness properties. Most solution concepts discussed in game theory impose as a condition the fact that a strategy profile must be an equilibrium, i.e. every player maximizes her expected utility at this strategy profile. As I have noted above, both Smit *et al.*'s and Guala and Hindriks' accounts also take this condition for granted in their discussion of institutions (the former implicitly, the latter explicitly). However, whatever the solution concept one wishes to retain to characterize institutions, it seems that this rationality condition is insufficient. This can be made quite transparent by using the example of the prisoner's dilemma (see Figure 3).

The prisoner's dilemma has a unique equilibrium, both Nash and correlated, corresponding to mutual defection. Indeed, it has been argued by some game theorists that the philosophers' persistent attempts to offer arguments for the rationality of cooperation in the prisoner's dilemma reveal a deep misunderstanding of the very nature and principles of game theory (e.g. Binmore, 2009). Consider however a Row player reasoning along the following lines:

> Given what I know and believe about the situation and the other player, I believe almost for sure that if I play D, Column will also play D. However, I also believe that if I play C, there is a significant chance that Column will play C.

https://doi.org/10.1017/S1744137418000073 Published online by Cambridge University Press

At first sight, it is not clear what is wrong with Row's conditional reasoning. Suppose that Row's conditional probabilities are p(Column plays D|I play D) = 1 and p(Column plays C|I play C) = ½. Then, Row's expected utilities are respectively $Eu(D) = 2$ and $Eu(C) = 3$. As a consequence, being Bayesian rational, Row should play C, i.e. should choose to play a dominated strategy. Is there anything wrong in Row reasoning this way? The definition of the correlated equilibrium solution concept excludes this kind of reasoning because, for any action s_i, the computation of expected utilities for each alternative action s_i' should be made using the conditional probabilities $p(.|s_i)$. In other words, it is implicitly assumed that the players' counterfactual beliefs should be consistent with a causal independence condition according to which one's strategy choice should have no causal influence upon others' strategy choice. More precisely, the players must believe that such a causal independence holds (Board, 2006). This can be made explicit in our semantic framework. For this, we need a semantic account of causal counterfactuals, which is obtained on the basis of a selection function f: $\Omega \times \Psi \rightarrow \Omega$ as introduced by Stalnaker (1968). This function states what is causally possible in any possible world x given some counterfactual event $[\phi]$. For any such possible world and counterfactual event, $f(x[\phi])$ corresponds to the set of possible worlds that would be possible were $[\phi]$ to occur in x. This selection function is a semantic counterpart to Lewis's system of spheres (see section A.3 of the appendix).

The kinds of causal counterfactuals that concern us are those where one considers deviating from what one is actually doing. To make sense of the use of counterfactual reasoning in this case, we need to add an assumption to the framework of section 2 stating that any player fully believes that she makes the strategy choice that she actually makes. This is formally stated as follows:

$$\text{For all } i \in N \text{ and all } x, y \in \Omega\text{: if } y \in B_i(x), \text{then } C_i(x) = C_i(y). \tag{8}$$

On this basis, it is now possible to formally express both the event that the players' strategy choices are causally independent and the event that a player i believes in the causal independence of the players' strategy choices (see section A.3 of the appendix). I call the latter condition (BCI). Denote $[bci]$ the set of worlds where condition (BCI) holds for all players. It follows from what has been said about the formal properties of the correlated equilibrium concept that for any game G where a correlated equilibrium is implemented, we have a *s.e.m.* $M = < \Omega, f, \{B_i, C_i, p_i\}_{i \in N} >$ satisfying Rationality, Common Belief of Rationality, Agreement but also Belief in Causal Independence as expressed by (BCI). Actually, since $[bci] = \Omega$, not only is each player believing in causal independence for strategy choices, but this is also common belief (Board, 2006).

This result provides a first reason to take into account counterfactual reasoning in the characterization of institutions, at least if the notion of correlated equilibrium is thought to be essential in this endeavor. But this also points out that the relevance of Guala and Hindriks's account of institutions depends

https://doi.org/10.1017/S1744137418000073 Published online by Cambridge University Press

on the status of the (BCI) condition. Indeed, as the use of the correlated equilibrium concept in their "rules-as-equilibrium" view makes the (BCI) condition mandatory, the latter must be thought as being constitutive of the nature of institutional practices. Whether or not this is the case is a difficult issue that I cannot discuss extensively here. It should be noted however that there exists a large documentation of past institutional practices that relied on what Jon Elster (2007) has characterized as "magical thinking" and that violate the (BCI) condition (or at least rest on dubious beliefs about the causal structure of the world). The concept of "self-confirming equilibrium" has been used to characterize what may look to an outside observer as "bizarre" behavior grounded on wrong beliefs (Fudenberg and Levine, 1993). Such behavior can be motivated by beliefs of causal *dependence* regarding strategy choices as illustrated by the example of the prisoner's dilemma above. Indeed, assume that Column reasons in the same way as Row. Then they will both cooperate, thus receiving a spurious confirmation of their initial beliefs. As long as none of the players experiments by defecting, they have no way of discovering the "wrongness" of their beliefs. However, they are rational by the standard of (some versions of) Bayesian decision theory and are even able to implement the Pareto-optimal outcome. A case can be made for the claim that causal independence of strategy choices is constitutive of the kinds of strategic interactions modeled by game theory. Rationality then seems to entail the players to taking into account what is deemed to be necessary.[12] This is however a normative claim made from the observer perspective that seems to have no bearing from the agent perspective.

6. Dynamic interactions and belief revision

Interestingly, the examples discussed by Smit *et al.* and by Guala and Hindriks are all examples of interactions corresponding to games in strategic form. That means that they are concerned with institutions regulating interactions where the players act either simultaneously or in ignorance of others' choices. Many relevant socioeconomic phenomena do not depend on these kinds of interactions, however. In fact, many institutions regulate interactions that are *sequential* and/or *repeated*. The players' reasoning in these kinds of interactions can then be grounded on information about others' past behavior. This has at least two major implications for any account of institutions grounded on a game-theoretic framework. First, the solution concept of correlated equilibrium is no longer the relevant one. Second, it becomes possible for players to receive information that contradicts one or several of their full beliefs. This will be especially the case when another player chooses an "out-of-the-equilibrium-path" strategy. It is then essential to provide an analysis of the way players *revise* their beliefs in the face of such surprising information. This section will focus on this second

12 This claim is forcefully made by Stalnaker (1996). For a dissenting view, see Levi (1997).

https://doi.org/10.1017/S1744137418000073 Published online by Cambridge University Press

issue as it emphasizes the role of counterfactual reasoning. For the sake of simplicity, I will simply assume that we are dealing with games of perfect information, though the whole discussion could easily be extended to games of imperfect/incomplete information. Indeed, it is important to note that rules and institutions are essential even in games of perfect information. As the analysis below will show, the players' beliefs at nodes that are out-of-the-equilibrium-path determine the range of parameter values (i.e. payoffs) for which a given strategy profile corresponds to a (subgame-perfect) equilibrium.

Now that we are dealing with dynamic interactions, a question that arises is whether our semantic framework is adapted to account for them. Indeed, the *s.e.m.* are models of *static* normal form games. There are two possibilities here. A first possibility consists in enriching the description of the strategic interactions by adding to the definition of G features specifying the game tree, especially information about the different nodes (decision nodes, chance nodes, terminal nodes). The semantic models of such a game should then specify what each player is actually doing and believing at each actual node as well as what they would do and believe at each node that is not actually reached. A second possibility is to embrace a notion of disposition. More precisely, following Stalnaker (1998: 40), we may interpret strategy choices as reflecting the players' *dispositions* to behave in specific ways in each possible situation.

This characterization of strategy choices imposes a consistency condition between a player's dispositions and what she would actually do were specific circumstances to arise. Assuming that the players are rational in the sense given above, this consistency depends on the beliefs they would have in these circumstances, i.e. what I have called "counterfactual beliefs." In other words, the study of dynamic games through semantic models imposes determining how the players revise their beliefs when they receive information, especially about others' behavior. Obviously, Bayesian conditioning will do for all the cases where the information corresponds to an event to which players have ascribed a strictly positive probability. But Bayesian conditioning is silent for "surprising" events, events that are inconsistent with the players' full beliefs and to which they initially ascribed a null probability. For these cases, we will have to define the players' counterfactual beliefs according to a belief revision selection function similar to the selection function of the preceding section. The kinds of surprising events that interest me here are those where a player has to make a decision at a node that she was not expecting to reach given others' behavior. In other words, we are concerned with the way a player *would* form beliefs about off-the-equilibrium path events arising because of someone else's deviation from the equilibrium. I will return below to the ontological and economic significance of these kinds of epistemic counterfactuals. Before that, I extend the formal framework to encompass the phenomenon of belief revision.

I have introduced above the notion of *objective* selection function f to reflect on the underlying causal structure of any institutional practice. However,

https://doi.org/10.1017/S1744137418000073 Published online by Cambridge University Press

belief revision concerns the way players subjectively change their perception of the world when they receive new information. We are thus interested in how individuals change their epistemic/doxastic attitudes about events, rather than what the causal features of the world are. This difference is captured by introducing a profile of *subjective* selection functions $\{f_i\}_{i\in N}$. The definition of each function f_i is given in section A.5 of the appendix.

If indeed the players believe in causal independence (condition (BCI) above), then that implies that active beliefs are irrelevant to the evaluation of the rationality of an action (Stalnaker, 1999). However, as I have noted above, there is no reason to take condition (BCI) as mandatory in a characterization of institutional practices. What about "passive beliefs," i.e. beliefs based on observation, evidence and inference, especially about others' behavior? Consider the following reasoning of a player named Ann who participates in an institutional practice along with Bob and Chris:

> If Bob were to choose strategy A, I (actually) believe that this would be causally irrelevant to the choice of Chris. However, this would lead me to believe it is very likely that Chris will also choose A rather than B or C. However, were Bob to choose B or C, I would have no particular belief regarding which choice Chris will make.

It seems to me that Ann's reasoning sounds perfectly reasonable. It can be rationalized for instance if we consider that the counterfactual supposition that Bob chooses A is articulated with the assumption that there may be a norm that makes choosing A mandatory. In this case, Ann's reasoning takes into account the fact that, even if she actually fully believes that the norm does not hold, some evidence might suggest the contrary. There is thus a clear distinction between (1) what Ann actually believes about the causal structure of the world, (2) what she actually believes about what others will do, and (3) what she would believe if she were to observe unexpected behavior from one or several other persons. This illustrates the crucial fact that while one may believe in causal independence when participating in an institutional practice, it is still reasonable to entertain a kind of *epistemic* dependence. Obviously, the objective selection function f cannot capture this distinction. But because in the case of passive beliefs there is no reason to expect counterfactual beliefs to be identical to actual beliefs, the subjective selection functions f_i can capture it. Since in the case of active beliefs the expression (A9) in the appendix indicates that the two kinds of functions are equivalent, there is no loss of generality by ignoring the function f in the characterization of an institution.

We now arrive at a point where we can fully characterize a game situation where a rule is followed as a tuple $\boldsymbol{G_\omega}$: $< G, < \Omega, \boldsymbol{\omega}, \{f_i, C_i, p_i\}_{i\in N} >>$. Correspondingly, the *ex ante* epistemic game $\boldsymbol{G}$: $< G, < \Omega, \{f_i, C_i, p_i\}_{i\in N} >>$ provides a formalization of an institution as a whole. The game situations contain a full description of the players' conditional reasoning, both evidential

https://doi.org/10.1017/S1744137418000073 Published online by Cambridge University Press

and counterfactual. As I have explained in the preceding sections, the players' counterfactual reasoning both characterizes their rationality and is constitutive of rule-following behavior. Clearly, belief revision and counterfactual beliefs are also dependent on counterfactual reasoning. It is worth noting that, from an economic perspective, the significance of counterfactual beliefs is not merely conceptual but also theoretical and empirical. A great example is offered by Avner Greif's (2006: chapter 9) comparative study of the organization of economic exchanges in two communities of traders during 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 interest 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 that poses the usual moral hazard problem. In spite of the fact that the two communities were using similar technologies, Greif argues (on the basis of historical archives) that they quite sensibly differed in the way they organized economic exchanges. In other words, they solved the commitment problem through two different sets of institutions. On the one hand, the Maghribi relied on a multilateral punishment strategy according to which an agent "cheating" a merchant would never be hired again by any merchant of the community. On the other hand, the Genoese used a one-sided punishment strategy where an agent's past behavior was not taken into account by the merchants in their decision to hire her. The cheated merchant would terminate the relationship but the chance of the cheating agent being hired by another merchant would be left unchanged. Both institutional arrangements solved the commitment problem but they led to a significant difference regarding the wages perceived by agents: Genoese agents used to receive higher wages than Maghribi agents because the former needed a higher wage premium to be deterred from cheating.

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" (Greif, 2006: 269–70). Cultural beliefs are directly responsible for equilibrium selection in a game because they provide focal points and help the coordination of expectations. They are thus 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, added emphasis). In the semantic framework that has been

developed throughout the paper, cultural beliefs are directly represented by the probability measures $p_{i,\omega}$ that characterize each game situation where a rule is followed. The most interesting feature is that cultural beliefs are defined explicitly for actions (nodes) both on and off the path of play. Regarding the former, they directly measure the probability that a given choice will actually be made under some circumstances. The latter are however unobservable: if the equilibrium play is indeed implemented, behavior off the path of play will never be observed. In these cases, cultural beliefs correspond to the *counterfactual beliefs* formalized above through the selection functions f_i and the probability measures $p_{i,x}(./.)$ (see section A.5 of the appendix).

The fact that cultural beliefs refer partially to unobservable events does not mean that they play no role in the institutional analysis. Quite the contrary, Greif suggests that they help to explain major differences between the two communities of traders that have had a long-lasting impact on the institutional trajectories of their economies. In particular, it seems that the Maghribi's "collectivist" cultural beliefs have directly constrained their ability to benefit from potential gain-from-trade opportunities with foreign merchants. Depending on the merchants' belief revision policy, it may or may not be worth engaging in inter-economy trade. Greif provides an analytical argument that Maghribi traders have been deterred from engaging in inter-economy trade because their cultural beliefs sustained a lower incentive-compatible wage to be paid to agents in the intra-economy trade. This was quite different for the Genoese merchants and their "individualistic" cultural beliefs. Their cultural beliefs relied less on information sharing than the Maghribi's traders, and though agents never actually cheated the off-the-path-of-play belief regarding the probability that cheaters would be re-hired was the same as for the honest agent. Engaging in inter-economy trade was thus intuitively less risky than for Maghribi's merchants, as the incentive-compatible wage paid to agents was already set on the supposition that cheating agents could be re-hired. The significant economic result was that the Genoese community quickly became integrated (thus benefiting from new gain-from-exchange opportunities), while the Maghribi community remained largely segregated. This case illustrates the point made above regarding the importance of rules even in perfect-information extensive-form games. The fact that agents do not hold the same counterfactual beliefs implies that the cooperative outcome is not an equilibrium among the Maghribis and the Genoese for the same payoff values. Higher wages were needed to foster cooperation among the Genoese than among the Maghribis.

I take this case study to provide a good illustration of the importance of counterfactual beliefs in institutional analysis. The way rational agents reason on the basis of counterfactual suppositions is thus essential both to characterize the nature of institutions but also to explain institutional mechanisms and trajectories regarding the organization of exchanges. This is especially significant in the perspective of a naturalistic social ontology that builds on the theoretical and empirical results of the social sciences.

7. Counterfactuals and the cognitive basis of rule-following behavior

This section provides a third reason for the importance of acknowledging counterfactual reasoning to achieve a better understanding of the nature of institutions. It is related to what can be called the cognitive basis of rule-following behavior. The two preceding sections have been interested in emphasizing the importance of causal and epistemic counterfactuals in the characterization of the equilibrium behavior that is constitutive of institutions under the rules-in-equilibrium view. The point has been to show, *assuming equilibrium play*, which kind of equilibrium holds and for which parameter values (especially payoffs) depend on the content of players' counterfactual beliefs and beliefs about counterfactual outcomes. A different but equally important issue concerns the way players are able to *converge toward equilibrium play*. Pointing out that such a convergence is achieved thanks to a rule is begging the question, because equilibrium play is part of the nature of any rule. What has to be determined is the cognitive mechanisms required to permit each player to reason their own way about what others are thinking and will be doing.

There exists a huge literature about learning in games that studies the learning mechanisms entailing convergence toward equilibrium play (e.g. Fudenberg and Levine, 1998). However, the issue at stake here is not how agents are learning a new rule (or equivalently, how new rules emerge), but rather how agents are able to achieve equilibrium play by reasoning, i.e. by making inferences regarding what they should do on the basis of what they know and believe. Such an ability is indeed constitutive of rule-following behavior and is related to a long-standing issue among philosophers and game theorists. An early discussion of this problem is of course to be found in the work of Thomas Schelling (1960), who emphasized the importance of 'focal points' to explain the ability of human agents to coordinate without communicating. According to Schelling, there are many features in a strategic interaction beyond the mathematical properties of a game that can be resources for individuals to help forming expectations about what others will do. Focal points may result from cultural, aesthetic or psychological factors that are sufficiently pervading for each person to consider that they are in one way or another 'obvious' to everyone else. David Lewis's (1969) theory of conventions and common knowledge similarly points out the importance of the salience of some outcomes in strategic interactions. Lewis more particularly insisted on the role played by the "force of the precedent." The point is that the fact that a particular outcome has obtained in previous interactions often gives reasons to believe that the same outcome will result in future similar interactions. According to Lewis's account, conventions originate and are maintained through the convergence of expectations that result from the salience of the precedent.

Without specific reference to a game-theoretic framework, the mechanisms underlying the "meeting of minds" required to solve coordination problems have

https://doi.org/10.1017/S1744137418000073 Published online by Cambridge University Press

been especially investigated by the "theory of mind" literature. The issue at stake is to determine how individuals are able to form convergent expectations about what each other will do and thus to act in accordance with these expectations. This corresponds to what is sometimes called "mindreading," i.e. the ability to read others' intentional attitudes by putting oneself into their shoes. Though there are several competing accounts of mindreading, the most dominant ones broadly fall within the scope of the so-called simulation theory (Goldman, 2008). Contrary to the main competing set of accounts (sometimes labeled the Theory-theory approach), the simulation theory does not assume that individuals are able to theorize about others' mental attitudes, but rather that they can project their own attitudes into others' minds. In other words, they use the product of their own reasoning as an indication of the output of others' reasoning.

An interesting discussion of mindreading in the context of simulation theory is offered by Morton (2003), who contrasts two approaches to the problem of the meeting of minds. The first is the *motive-based* approach, where agents "think out the motives and reasoning of each agent, and try to predict what decisions will result, as part of thinking out what would be best for her herself to do" (Morton, 2003: 18). The second is the *solution-based* approach, where agents "try to define some equilibrium outcome, focusing on the properties that a situation would have if each person were reacting to the motives of each other but not directly representing the motives, and extract both expectations and decisions from it" (Morton, 2003: 18). The former has two main disadvantages compared to the latter. On the one hand, it is more complex to use because it requires to form nested intentional attitudes of the kind "I believe that you believe that I believe ..." On the other hand, it may not produce a definite practical answer regarding what one should do, especially in situations where multiple equilibria exist. According to Morton (2003), the solution-based approach faces neither of these difficulties. Schelling's focal points and Lewis's force of the precedent are essentially versions of this latter approach as they build on the individuals' ability to single out one specific outcome from the range of possible ones on the basis of some criterion and then to assume that everyone else is seeing things in the same way. As noted by Guala (2016: 97), such solution-thinking leads to something similar to Lewis's assumption of symmetric reasoning according to which each individual reasons in such a way that she attributes to others the same inferences that she is making herself.

The issue of how individuals are able to converge toward the same solution remains unanswered yet. As the term "simulation theory" indicates, this occurs through a process where each individual simulates others' thinking by taking for granted that whatever is obvious for her is also obvious for others. What is especially relevant is the way such a simulation is made. Morton (2003: 122–6) distinguishes two kinds of simulation that he calls "cocognition" and "modeling." Crucially, both kinds of simulation appeal to conditional reasoning with modeling depending on the ability to engage in *counterfactual* reasoning:

https://doi.org/10.1017/S1744137418000073 Published online by Cambridge University Press

"Being able to think in subjunctive or counterfactual terms what a person or group would do increases one's capacity to assess the salience of outcomes in a coordination problem" (Morton, 2003: 139). The point is that making use of the modeling form of simulation requires the ability to imagine what I *would* do if I *were* in the situation of the person I am trying to coordinate with. This thinking process greatly enhances our capacity to identify salient outcomes and thus to work out practical conclusions about what we should do.[13]

It might be useful to finish by considering how all of this is relevant in accounting for the cognitive basis of rule-following behavior. In the two previous sections, I have distinguished two kinds of counterfactuals that are essential to characterize institutions and rule-following behavior in the rules-in-equilibrium view. Causal counterfactuals concern what would happen if a player were to deviate from the equilibrium play. Epistemic counterfactuals correspond to the beliefs the players would have if a player were to deviate from the equilibrium play. I have argued that both must enter into the characterization of a rule because a change in either of them may entail a change in the actual behavioral pattern or a change in the conditions under which a given behavioral pattern will indeed correspond to an equilibrium. It is obvious that speaking of causal and epistemic counterfactuals in this sense implies that rule-followers are actually able to engage in counterfactual reasoning. What this section further establishes is that this ability may itself have arisen from the necessity to solve coordination problems on the basis of the identification of focal points. If indeed the meeting of minds happens mostly through forms of simulation thinking like modeling, then individuals must be able to form expectations by imagining what they would think and do if they were in someone else's situation. In itself, this is already a special form of counterfactual belief that is used as a basis to infer others' actual beliefs and actions. To put it another way, counterfactuals are essential in characterizing an institution in a somewhat static fashion. They are also central to explaining how institutions may have arisen in the first place.

8. Conclusion

My purpose in this paper has been to pursue Smit *et al.*'s and especially Guala and Hindriks's naturalistic accounts of social ontology. I agree with these authors that a broad game-theoretic framework, and more specifically the concept of correlated equilibrium, are the appropriate starting points for dealing with the issue of the nature of institutions. However, I have also pointed out the necessity to go further in considering explicitly the way players reason when

13 A referee has pointed out that another realm of strategic reasoning that requires counterfactual reasoning is "Machiavellian reasoning." Machiavellian interactions are exploitive rather than cooperative. According to the so-called "Machiavellian intelligence hypothesis," the large brain of humans evolved from the intense social competition for reproduction (Jackson, 2012). The hypothesis is controversial, but its truth would only strengthen the point made in this section.

they participate in an institutional practice. I have particularly emphasized the importance of the players' conditional reasoning, which is based on the supposition that some factual or counterfactual event holds. I have argued that the best way to deal with these issues is through semantic models satisfying several properties. The key lesson is that the functioning of institutions depends both on the players' beliefs about causal counterfactuals and the players' counterfactual beliefs about others' behavior. This shows that more attention should be given to features that do not transpire directly in people's behavior. In some way, the whole discussion can thus be seen as a broad argument against behaviorism in philosophy and in the social sciences.

Acknowledgments

This paper has been presented at Les Treilles III workshop "Coordination in Economics" (May 29–June 3, 2017). I thank the participants of this workshop for their valuable comments. I also thank three anonymous referees for their challenging remarks. All errors and omissions are mine.

References

Aumann, R. (1987), 'Correlated Equilibrium as an Expression of Bayesian Rationality', *Econometrica*, **55**(1): 1–18.

Aumann, R. and J. Dreze (2008), 'Rational Expectations in Games', *American Economic Review*, **98**(1): 72–86.

Bacharach, M. (1993), 'When Do We Have Information Partition?' in M. Bacharach, M. A. H. Dempster, and J. Enos (eds), *Mathematical Models in Economics*, Oxford: University of Oxford, pp. 1–23.

Bennett, J. (2003), *A Philosophical Guide to Conditionals*, Oxford: Oxford University Press.

Binmore, K. (2009), *Rational Decisions*, NY: Princeton University Press.

Board, O. (2006), 'The Equivalence of Bayes and Causal Rationality in Games', *Theory and Decision*, **61**(1): 1–19.

Bradley, R. (2016), *Decision Theory with a Human Face*, Mimeo, London School of Economics.

Brandenburger, A. (2014), *The Language of Game Theory: Putting Epistemics into the Mathematics of Games*, London: World Scientific.

Dietrich, F. and C. List (2016), 'Mentalism versus Behaviourism in Economics: A Philosophy-of-Science Perspective', *Economics and Philosophy*, **32**(2): 249–81.

Elster, J. (2007), *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*, Cambridge: Cambridge University Press.

Fudenberg, D. and D. Levine (1993), 'Self-Confirming Equilibrium', *Econometrica*, **61**(3): 523–45.

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

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

Goldman, A. (2008), *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*, Oxford, New York: Oxford University Press.
Greif, A. (2006), *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade*, Cambridge: Cambridge University Press.
Guala, F. (2007), 'The Philosophy of Social Science: Metaphysical and Empirical', *Philosophy Compass*, **2**(6): 954–80.
Guala, F. (2016), *Understanding Institutions: The Science and Philosophy of Living Together*, New York: Princeton University Press.
Guala, F. and F. Hindriks (2015), 'A Unified Social Ontology', *Philosophical Quarterly*, **65**(259): 177–201.
Gul, F. B. and W. Pesendorfer (2008), 'The Case for Mindless Economics', in A. Caplin and A. Schotter (eds), *The Foundations of Positive and Normative Economics*, Oxford: Oxford University Press, pp. 3–39.
Hands, D. W. (2013), 'Foundations of Contemporary Revealed Preference Theory', *Erkenntnis*, **78**(5): 1081–108.
Hédoin, C. (2017), 'Institutions, Rule-Following and Game Theory', *Economics and Philosophy*, **33**(1): 43–72.
Hindriks, F. and F. Guala (2015), 'Institutions, Rules, and Equilibria: A Unified Theory', *Journal of Institutional Economics*, **11**(3): 459–80.
Jackson, M. (2012), 'Machiavellian Intelligence Hypothesis', in N. M. Seel (eds), *Encyclopedia of the Sciences of Learning*, Boston, MA: Springer.
Levi, I. (1997), *The Covenant of Reason*, Cambridge: Cambridge University Press.
Lewis, D. (1969), *Convention: A Philosophical Study*, Oxford: Blackwell Publishing.
Lewis, D. (1973), *Counterfactuals*, Oxford: Blackwell Publishing.
List, C. and, K. Spiekermann (2013), 'Methodological Individualism and Holism in Political Science: A Reconciliation', *American Political Science Review*, **107**(4): 629–43.
Morton, A. (2003), *The Importance of Being Understood: Folk Psychology as Ethics*, London: Routledge.
Perea, A. (2012), *Epistemic Game Theory: Reasoning and Choice*, Cambridge: Cambridge University Press.
Ross, D. (2014), *Philosophy of Economics*, New York: Palgrave Macmillan.
Schelling, T. (1960), *The Strategy of Conflict*, Harvard: Harvard University Press.
Sillari, G. (2013), 'Rule-Following as Coordination: A Game-Theoretic Approach', *Synthese*, **190**(5): 871–90.
Smit, J. P., F. Buekens and S. du Plessis (2011), 'What Is Money? An Alternative to Searle's Institutional Facts', *Economics and Philosophy*, **27**(1): 1–22.
Smit, J. P., F. Buekens and S. du Plessis (2014), 'Developing the Incentivized Action View of Institutional Reality', *Synthese*, **191**(8): 1813–30.
Smit, J. P., F. Buekens and S. du Plessis (2016), 'Cigarettes, Dollars and Bitcoins – An Essay on the Ontology of Money', *Journal of Institutional Economics*, **12**(2): 327–47.
Stalnaker, R. (1968), 'A Theory of Conditionals', in W. L. Harper, R. Stalnaker and G. Pearce (eds), *IFS*, University of Western Ontario Series in Philosophy of Science 15, Springer Netherlands, pp. 41–55.
Stalnaker, R. (1996), 'Knowledge, Belief and Counterfactual Reasoning in Games', *Economics and Philosophy*, **12**(2): 133–63.
Stalnaker, R. (1998), 'Belief Revision in Games: Forward and Backward Induction', *Mathematical Social Sciences*, **36**(1): 31–56.

Stalnaker, R. (1999), 'Extensive and Strategic Forms: Games and Models for Games', *Research in Economics*, **53**(3): 293–319.
Tomasello, M. (2014), *A Natural History of Human Thinking*, Harvard: Harvard University Press.
Wittgenstein, L. (2010), *Philosophical Investigations*, Oxford: Blackwell Publishing.