Conflation of agency models in Francesco Guala's social ontology sheds light on how normativity of social institutions might have evolved

Valerii Shevchenko

HSE University, Moscow, Russian Federation
https://orcid.org/0000-0001-5040-8897

Рекомендация для цитирования:
Shevchenko V. (2026) Conflation of Agency Models in Francesco Guala's Social Ontology Sheds Light on How Normativity of Social Institutions Might Have Evolved. *Социология власти*, 38 (2): 187-229.
EDN: UXJNAO

For citation:
Shevchenko V. (2026) Conflation of Agency Models in Francesco Guala's Social Ontology Sheds Light on How Normativity of Social Institutions Might Have Evolved. *Sociology of Power*, 38 (2): 187-229.

Поступила в редакцию: 05.08.2025; прошла рецензирование: 17.12.2025; принята в печать: 11.01.2026
Received: 05.08.2025; Revised: 17.12.2025; Accepted: 11.01.2026

© Author, 2026
This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract: According to Hannah Arendt, a hallmark of the modern era is the alienation of the individual from the world, understood, in part, as a loss of interest in the common things that mediate political process. One of the origins of modern alienation for Arendt was the discovery of the New World and the ensuing explosion of cartography. This article attempts to demonstrate the political potential of maps as a means of "re-engaging with the world". Following Bonnie Honig, one could argue that the pre-modern tradition of participatory cartography has such potential. The reconnection proposed by this tradition could lie in the transformation of the world reflected on the participatory map into a thing that can serve as a subject of discussion and a basis for the self-identification of the collective. As a positive illustration of the work of participatory cartography today, the article turns to the late work of Bruno Latour. Seeing modern cartography as a means of reproducing the erroneous notion of the Earth as a closed totality of the Globe, Latour proposes as an alternative other principles of cartography appropriate to the age of the new climatic regime. Some of these principles, in particular the identification of intangible and not always visible agents of territory on cartographic projections, formed the basis of his last experiment 'Où Atterrir?'. In it, the final maps of territories produced by the research participants become a means of making the public's concerns visible, ensuring the sustainability and accessibility of the political space.

Keywords: social ontology, rules-in-equilibria, correlated equilibrium, constitutive rules, social institutions, normativity, game theory

Что смешение моделей агентности в социальной онтологии Франческо Гуала говорит об эволюции нормативности социальных институтов

Валерий Шевченко

Национальный исследовательский университет «Высшая школа экономики», Москва, Российская Федерация
https://orcid.org/0000-0001-5040-8897

Резюме: В статье критически анализируется унифицированная социальная онтология Франческо Гуалы, а именно теория «правил-в-равновесии» (RiE), разработанная совместно с Фрэнком Хиндриксом. RiE пытается примирить объяснения социальных институтов на основе правил и равновесий, утверждая, что конститутивные правила Джона Сёрла — лишь лингвистические трансформации регулятивных правил, образующих коррелированные равновесия в координационных играх. Несмотря на новаторский подход, теория содержит фундаментальные проблемы, подрывающие её объяснительную адекватность. Во-первых, выявляется конфляция моделей агентности в RiE. Гуала отождествляет паттерны решения стратегических проблем у животных (как эволюционно стабильные стратегии) с коррелированными равновесиями в решениях людей, несмотря на их расходящиеся допущения: первые являются онтическими, вторые — эпистемическими и байесовскими. Эта конфляция проистекает из неверного применения концепции «буржуа-равновесия» Джона Мэйнарда Смита, описывающего эволюционную стабильность, а не коррелированные стратегии. Во-вторых, обнаруживается критическая двусмысленность понятия «репрезентация» в RiE. Хотя Гуала обосновывает уникальность людей через когнитивную способность к расщеплению стимула и реакции, теория не объясняет механистически, как эта способность порождает нормативные правила из структур равновесий. Репрезентация функционирует одновременно как отношение между концептами и вещами и как когнитивная способность живых организмов, но их связь остается неразработанной. В итоге нормативность вводится в RiE без адекватного объяснения. Статья предлагает переход от репрезентационной к онтической объяснительной модели, опираясь на теорию конструирования ниши для объяснения того, как способности к расщеплению стимула и реакции позволяют нормативным социальным институтам возникать из «животных конвенций».

Ключевые слова: социальная онтология, правила-в-равновесии, коррелированное равновесие, конститутивные правила, социальные институты, нормативность, теория игр

Introduction

Social conventions are of major interest for social ontology, a philosophical study of the entities comprising the social world. Some scholars see them as foundational blocks of social ontology (Bickhard, 2008) and others use them as more primitive units to construct complex social ontologies involving the notions of *social institutions*, which are often dubbed normatively-driven sets of social conventions (Aoki, 2007; Hindriks, 2022). For example, Guala (2016) mentions that such mundane notions as "a professor", "a student" or "a husband" make no sense without social institutions as normative structures embedding and making these roles and their content possible. It is because these notions are a part of a mostly hidden and taken-for-granted infrastructure of roles with duties and rights. And conventions, as Guala argues, hold those together.

A good deal of the current social ontology is detached from actual social science (Sarkia & Kaidesoja, 2023). It has mostly been developed within analytic philosophy (Baker, 2019; Epstein, 2015; Gilbert, 1992; Searle, 1995) and social scientific tradition of critical realism (Archer, 1995; Elder-Vass, 2010). There are views that social ontology is intrinsically more complex than any "natural" ontology for being "interactive", or dependent on beliefs of individuals and hence changeable (Epstein, 2015; Hacking, 1999) and that social reality is inherently normative, meaning it cannot be descriptive and hence cannot import methodologies from "natural" ontologies (Ásta, 2024; S. A. Haslanger, 2012). Such a position renders and justifies the importance of an "extra-scientific" study of the social world. To me, both interactivity and inherent normativity of social reality do not prevent their explanation from a naturalistic angle, which could shed light on a descriptive social ontology.

According to Epstein (2018), the field of social ontology divides into two distinct inquiries. The first one deals with the constituents of social entities and addresses the question "what is the social world composed of?". The second strand of research is concerned with the construction of social categories, or kinds, and with the question "how do social kinds like money, borders, marriage and others get established?". Individual people constituting a social group exemplify the former inquiry and children playing a game where stuffed animals have a tea party exemplify the latter (as they have "set up" the conditions for something to be a tea party) (Epstein, 2015, p. 57).

The difference between the strands is in the metaphysical relation of individuals and social facts. In the first case, social facts supervene on the facts about individuals, meaning that social properties cannot change without changing the individual ones. In the second case, facts about individuals *set up the conditions* for something to count as a social

fact, i.e., dollar bills with a particular ink and paper and collective acceptance for money (2015, p. 58).

One of the most famous expressions of the latter is Searle's (1995) formula "X counts as Y in context C". For example, "bills issued by the Bureau of Engraving and Printing (X) count as dollars (Y) in the United States (C)" (1995, p. 28). Epstein calls the former relationship (exemplified with "issued by") *grounding* and the latter one (exemplified with "counts as") *anchoring*, where grounding is responsible for *instantiation conditions* for being, for instance, money, in a particular context, and anchoring is responsible for *the mechanism of establishing instantiation conditions*. For example, in Searle's case, *collective acceptance* of the fact X is such a mechanism[1].

Such consideration exemplifies one of the two main views on the role of metaphysics in scientific inquiry. The first, empiricist one, renders metaphysics irrelevant and redundant or, if taken seriously, expects its constraints on models to impede the knowledge growth. The other view presupposes special metaphysical concepts beyond the vocabulary of science itself. For example, they discern grades of possibility and necessity largely unknown to science and go beyond causal dependence widely used in science (Ross, 2023).

Ross (2023) suggests the distinction of *analytic* and *scientific* social ontology, where the former is derived from conceptual analyses and the latter is based on empirical investigations. For Ross, 'naturalistic metaphysics' as he calls it, does not apply transcendental concepts except those featured in theories, models or explanations of "first-order science". Ross defends what he calls 'radically naturalistic metaphysics' which is aimed at promoting scientific progress and might involve unifying or synthesising scientific discoveries.

Physics provides a vivid contrast for comparing scientific and analytic metaphysics. Ontology of physics is formulated in mathematical terms requiring some technical proficiency. Formulating ontology of physics in an analytic way with "possible worlds" and "propositions" would be counterintuitive and unnecessarily difficult. It would require translation of already existing scientific concepts into the language of analytic metaphysics.

Social science and social ontology do not enjoy straightforwardness of mathematical formulations of its concepts like physics does[2]. Instead, it invites "folk" concepts like "beliefs", "groups" and "social facts". Further

1 Searle's formula involves both relations, grounding and anchoring, but for the sake of simplicity and to illuminate collective acceptance as a necessary condition for Searle's formula I will refer to it as to anchoring.

2 Although not widely known, there have been such attempts. See Fararo (1989) and, most notably, Gintis (2007a).

divide comes from the attitude towards this ambiguity. Some scholars argue that such "folk" ontologies are sufficient Thomasson (2003) and others put an interplay of "folk" and social scientific ontologies to the fore by bridging what Sellars (1962) called "manifest" and "scientific" image of phenomena within their social ontological theorizing (Guala, 2016; S. Haslanger, 2018). The ambiguity between "folk" and scientific social ontology might indeed depend on mathematical apparatus and its underdeveloped character in social science. Ross (2023) also notes that ontology construction most probably will require its formulation with mathematical frameworks sensitive to structural modeling (like hierarchical Bayesian inference[1]) and not set theory and similar tools from analytic metaphysics.

As already mentioned, analytic social ontology is based on conceptual analysis and claims that it is prior to social scientific methodology (Epstein, 2016; Lauer, 2019; Searle, 1995). The naturalistic camp criticizes this approach as detached from actual social science (Elder-Vass, 2007; Kincaid, 2024; Little, 2020; Ross, 2023; Sarkia & Kaidesoja, 2023). Their focus is meta-ontological rather than ontological as they wonder what it means to ask questions about ontology of social science and what are the ways to arrive at an ontology instead of proposing or criticizing particular ontological pictures.

Our main focus will be on Guala's "unified social ontology" (Guala, 2016; Guala & Hindriks, 2015; Hindriks & Guala, 2015) as it tries to save the "best of both worlds" from analytic and naturalistic social ontology by providing a scientifically rigorous treatment of classic analytic social ontology of Searle (1995).

In its standard formulations, which Guala (2007) calls the "Standard Model of Social Ontology" (SMOSO), social ontology describes the loosely constrained individualistic foundations of social phenomena and has three key elements (Tuomela, 2002):

- reflexivity
- performativity
- collective intentionality.

Reflexivity is a property of social entities to be largely comprised of beliefs about beliefs. There are I-mode and we-mode formulations of reflexive beliefs. Some philosophers say that initial and most basic beliefs comprising "the fabric" of the social are essentially in We-mode and are not reducible to I-mode (Gilbert, 1992; Schmid, 2023; Tuomela, 2002). However, there are also more individualistic accounts of reflexive beliefs based on game theory (Bicchieri, 2005; Guala, 2016).

1 Hierarchical Bayes is a model with probabilistic parameters (meaning they are uncertain) which allows for hierarchical relations between random variables.

Performativity amounts to social entities needing to be continuously maintained, performed or recreated. And collective intentionality, in its turn, refers to joint directedness of multiple individuals towards a phenomenon that contributes to its constitution. Collective intentionality tends to be presented either as a derivative of common knowledge and I-beliefs of the form "everyone knows that everyone knows that P", where P is some social fact like social norm (Bicchieri, 2005), or as a primitive notion which makes common knowledge redundant. Moreover, there are attempts to naturalize collective intentionality by showing its irreducibility to individual intentionality (Gallotti, 2012; Rakoczy & Tomasello, 2007).

A prominent example is Searle (1995) who asks whether it is possible to be epistemologically objective about ontologically subjective issues. How can we know the truths about things whose existence depends on our representations or feelings, for example, about money, property and marriage? By analysing these distinctions of ontology/epistemology and objectivity/subjectivity, Searle arrived at an idea of a missing ingredient that allows for a picture of ontologically subjective entities, which is *constitutive rules* of the form "X counts as Y in C".

Here, our classifications of the social world help establish and maintain it, whereas non-social objects are indifferent to our classifications of it, as Hacking (1999) puts forward with his distinction of interactive and indifferent kinds[1]. Nature's objects do not change their behaviour given these classifications of them as opposed to social objects. This idea illustrates the notions of reflexivity and performativity characteristic for the "Standard model". If social entities are comprised of beliefs about beliefs, their nature depends on these beliefs, and if beliefs change, social entities change accordingly. If social entities depend on beliefs about them, it is needed to constantly perform those to maintain them. To do this, individuals need to have collective intentionality about these beliefs. For example, for money to be itself, a relevant community holds a collective intention to believe that certain physical entities can be used as a medium of exchange.

Along with the "Standard model", Guala (2007) calls game theory one of the two main approaches to social ontology. He departs from the works of Schelling (1980) and Lewis (1969/2008), claiming that the latter was the first to apply rational choice analysis to social ontology and to present social conventions as systems of self-sustaining beliefs. Schelling, before Lewis, was already aware that coordination

1 This distinction mirrors the classic one of natural and social kinds, where the former are "homeostatic property clusters", sets of necessary and stable features (Boyd, 1991).

problems are unsolvable from the pure game-theoretic perspective if the existence of focal points is admitted, for the latter need a separate explanation. Schelling advocated the supplementation of an a priori approach to game theory with empirical studies of the emergence of focal points. This point was further explicated by Sugden & Zamarrón (2006).

An intrinsic problem of application of game theory to social ontology is that there is no mechanism of spreading of high utility-bearing strategies across the population. Seen in terms of biological fitness, the most successful strategy derived from a long evolutionary history spreads across the population. In contrast, seen from the utility viewpoint, the very cause of spread a utility-bearing strategy begs an explanation. This is why Guala critiques application of game theory to social ontology and tries to modify it.

Guala strives to overcome the shortcomings of both "Standard model" with its utter anti-realism (due to dependence on beliefs) and game theory in his own "unified social ontology", which integrates the notions of reflexivity, performativity and intentionality with formal accounts of social conventions as objective structures underpinning the aforementioned features of the "Standard model".

Hindriks & Guala (2015) propose a theory that social institutions like money, marriage or others can be seen as (sets of) conventions in Lewisian sense with added normativity.

Guala's and Hindriks's "rules-in-equilibria" theory

Hindriks & Guala (2015) present a unified theory of social institutions as rules in equilibria (RiE) represented symbolically by theoretical terms like "money" or "marriage". It bridges accounts of regulative rules, equilibria of strategic games and Searlian constitutive rules, where the former two are complementary and comprise a rules-in-equilibria account, and the latter supplements it by providing a symbolic representation. They essentially bridge Searle's account of institutional facts as rules with Lewis's theory of conventions seen through the lens of Aumann/Vanderschraaf/Gintis (Aumann, 1987; Gintis, 2009; Vanderschraaf, 1998). They reduce constitutive rules to regulative ones (Hindriks, 2005), which they represent formally as conditional strategies "if X, do Y" (Guala, 2016). To better understand their proposition, let us start from Searle's constitutive-rule theory of institutions.

Searle's social ontology distinguishes two kinds of rules: *regulative* rules, which govern actions that can occur independently, and *constitutive* rules, which create new kinds of social reality (Searle, 1995, 2010). In Searle's formulation, constitutive rules take the schematic form:

X counts as *Y* in context *C*

where *X* is a pre-institutional entity or action, *Y* is a *status function*, a social role or function assigned to *X*, and *C* is the relevant context or domain (Searle, 1995). For example, "putting the ball into the net (*X*) counts as scoring a goal (*Y*) in a game of football (*C*)" (Searle, 1995). Such rules do not merely regulate pre-existing behavior; they *create* new social facts. In Searle's own words, "institutional facts only exist within systems of constitutive rules" (Searle, 1995).

- **Constitutive vs. Regulative.** A constitutive rule makes a novel institutional action possible, whereas a regulative rule simply prescribes behavior within an already existing framework (Searle, 1995). Chess provides a classic example: the constitutive rules of chess *create* the possibility of the game, whereas a regulative rule would say, for instance, "if you touch a piece you must move it" (Searle, 1995).
- **Status Functions and Deontic Powers.** Under a rule *X* counts as *Y* in *C*, *Y* is a *status function* attached to *X*, and carrying this status typically confers normative powers (rights, obligations, etc.) on the bearer. Thus if a community collectively accepts that certain actions or objects bear status *Y*, those actions have deontic powers. Searle often emphasizes that constitutive rules imply deontic powers: e.g. a wedding ring (*X*) gives someone the status of "married person" (*Y*), along with associated rights and duties. In Searle's framework, linguistic declarations often play a role: he introduces the idea of a *Status Function Declaration*, a speech act that imposes or announces status functions as binding (Searle, 2010).

Searle identifies *institutions* with systems of constitutive rules. He writes that "an institution is any system of constitutive rules of the form *X* counts as *Y* in *C*" (Searle, 1995). Thus, for Searle, political offices, legal entities, money, marriages, etc., exist because underlying constitutive rules assign new status functions to physical or social substrates. These rules are held in place by *collective acceptance* of the community. In Searle's view, the syntax "*X* counts as *Y* in *C*" — often called the *counts-as* locution — succinctly captures the logic of institutional facts.

Searle's theory posits that social reality is built upon constitutive rules that are *creatively* implemented: they not only *regulate* behavior but *generate* the very phenomena like institutions and roles they describe (Searle, 1995, 2010).

Hindriks (2005) has challenged several aspects of Searle's constitutive-rule framework. His deconstruction focuses on the notions of *status functions* and the role of language. Broadly, Hindriks argued that Searle's theoretical apparatus is misleading, and that a more streamlined ac-

count can be given by focusing on collective acceptance and normative powers.

- **Status Functions as Deontic Powers.** Hindriks finds the term *status function* confusing and somewhat redundant. He suggests dropping the "function" and simply treating statuses as normative powers. In his words, "we can do without the term function while retaining the term status," instead explicating statuses directly as the bundle of deontic powers they grant (Hindriks, 2009). By equating statuses with deontic powers (rights, obligations), Hindriks makes the normative dimension of institutions explicit, rather than hiding it under the metaphor of a "function" (Hindriks, 2009). Indeed, Searle himself has acknowledged that "all status functions are deontic powers," which supports Hindriks's move toward a more direct terminology (Hindriks, 2009).
- **Redundancy of Status Function Declarations.** Searle's idea of a *Status Function Declaration* — a speech act that supposedly creates or recognizes a status — is, for Hindriks, unnecessary. He argues that the two key claims Searle attributes to such declarations (that collective acceptance is necessary and sufficient for the status) are already implicit in the standard "counts-as" formulation. Once we accept that *institutional statuses require collective acceptance, and that collective acceptance alone brings them into being* (the "Collective Acceptance Principle"), the special notion of a Status Function Declaration adds nothing new (Hindriks & Guala, 2015). Introducing declarations suggests without argument that only explicit speech acts can create institutions; Hindriks finds this unjustified and unhelpful. He concludes that Searle's extra machinery (the Status Function Declaration with its "double direction of fit") should be abandoned since it "does not add anything of value" (Hindriks & Guala, 2015).
- **Linguistic vs. Normative Distinction.** Hindriks also questions Searle's emphasis on language as the source of all institutional power. In his earlier work, Hindriks has argued that the regulative/constitutive distinction is mainly a grammatical one: regulative rules are phrased with explicit imperatives or deontic terms, while constitutive rules are phrased with the "counts-as" locution, but both embed the same normative content (Hindriks, 2009). Normative obligations figure explicitly in regulative rules ("Do X" / "If Y do X"), whereas constitutive rules imply those obligations without stating them overtly. Thus, the locus of normativity is not really different between the two; only the linguistic presentation is. Hindriks calls for a view of institutions that centers on *collective commitment* and acceptance of standards, rather than on linguistic declarations per se (Hindriks, 2009).

In sum, Hindriks dismantles Searle's superstructure of status functions and declarations, proposing instead that we should "explicate statuses in terms of normative powers" and rely on a simpler collective-acceptance principle (Hindriks & Guala, 2015; Hindriks, 2009). On this view, institutions are upheld by groups collectively endorsing certain rules, and the resulting normative powers of those rules are what really matters. His critique paves the way for unifying Searle's approach with more analytical models, by translating constitutive claims into the language of regulative rules and equilibria.

Guala & Hindriks (2015) present a synthesis of rule-based and game-theoretic views of institutions. They argue that constitutive-rule accounts and equilibrium-based accounts common in economics can be integrated using the concept of CE. Their core idea is that *constitutive rules are not ontologically fundamental*, but can be reconstructed from systems of regulative rules under coordination equilibria in iterated games (Guala & Hindriks, 2015).

As we mentioned, in a CE, each agent effectively follows a *conditional strategy* of the form "if *X* do *Y*." Guala & Hindriks note that this is just a regulative rule (Guala & Hindriks, 2015). For example, two herders might adopt strategies "graze if river is North" vs. "graze if river is South," thus solving a coordination problem. Each agent's strategy is equivalent to a rule prescribing what to do in a given circumstance. Thus, institutions based on coordination can be viewed as *collections of regulative rules* that form a stable equilibrium.

Moreover, the familiar constitutive formulation can be derived as a shorthand. Guala & Hindriks show that by introducing new institutional terms one can transform regulative conditionals into "counts-as" form. This two-part statement has a consequent identical to an institutional status. Compressing these, one obtains a constitutive rule:

A piece of land north of the river (*X*) counts as Nuer property (*Y*)
in the context of location (*C*)

In this way, Guala & Hindriks interpret "*X* counts as *Y*" rules merely as economy of description: they package together a base rule (antecedent) and a status rule (consequent) that were already implied by the equilibrium of regulative rules (Guala & Hindriks, 2015). They argue that any regulative rule can be converted into a constitutive rule by inserting an institutional term, and conversely any constitutive rule can be expanded back into regulative form. The purported novelty of constitutive rules is thus secondary: they label what is already established by agents' correlated strategies. Institutions "consist of regulative rules" from which constitutive formulations can be omitted without loss.

Guala's unified theory posits that institutions emerge from solving repeated coordination problems: agents arrive at a CE through a coordination device or focal point. The equilibria are characterized by mutual conditional strategies "if *X* do *Y*". In this picture, the content of any constitutive rule comes down to a cluster of conditional incentives and conventions that are already present in the equilibrium. Institutional terms like "dollar" or "married" are introduced for shorthand, but Guala emphasizes they are only instruments of "cognitive economy" and *do not possess any independent causality* (Guala & Hindriks, 2015).

Crucially, Guala reframes Searle's project in game-theoretic terms without surrendering its insights. He acknowledges that institutional statuses influence how we classify and act, but maintains that these statuses are always aligned with behavioral regularities expressed as CE. Thus Searle's emphasis on "counts-as" locution can be recovered as an epiphenomenon of coordination: an apparent creative *power of language*[1], but in fact nothing more magical than the effect of correlated strategy profiles. As Guala & Hindriks put it:

> "language is one among many coordination devices, and has no more creative power than a coin toss or any other event the players may use to coordinate their decisions" (Guala & Hindriks, 2015).

A key task is to explain normative powers under this account. Guala & Hindriks argue that normative aspects can be modeled as payoff-modifications of the underlying game akin to δ-parameters of Crawford & Ostrom (1995). A normative rule adds incentives or penalties that make certain actions like cooperation more attractive. In practice, this means

1 Searle makes deontic powers of institutional facts dependent on language. At the same time, evidence from cognitive archaeology suggests that early hominins were able to solve coordination problems without language. Sterelny (2021) argues that social institutions (or the human social contract) conceived as shared, enforceable norms regulating cooperation and reciprocity emerged *before the advent of complex language*. Archaeologically, early hominins engaged in cooperative foraging and coordinated tool use, such as the collective extraction of high-value resources, indicating mutualistic collaboration and implicit norm enforcement in small mobile bands from around 1.8 million years ago (Sterelny, 2016; Sterelny, 2021). Comparative primate research supports the plausibility of *pre-linguistic cooperation and rudimentary norm enforcement*, highlighting that great apes exhibit forms of mutualism and social regulation that likely scaffolded early human cooperation (Birch, 2017). Together, these findings support Sterelny's model that proto-social contracts rooted in norms, reputation, and coordination were present in pre-linguistic societies, with language later enabling more abstract, scalable, and formalized norms as human social groups expanded. Although Searle did not conceive of institutions as of solutions to coordination problem, this is relevant from a naturalistic social ontology point of view.

that if agents gain a "right" or incur an "obligation," we represent this by inserting costs or benefits into the payoff matrix. This transforms a general-sum game into a coordination game where the efficient equilibrium becomes more salient.

Guala shows that adding such normative costs can create new equilibria that were not present before, or make the socially optimal outcome stable. Importantly, these modifications do not require a distinct ontological category beyond standard game-theoretic tools. Normative powers are simply part of the equilibrium framework: they enable coordination by altering incentives[1]. Guala & Hindriks illustrate that any status rule such as rights to use, transfer, or exclude can be recast as regulative rule once normative powers are included.

In effect, Guala & Hindriks endorse a transformation view: any institution describable by a constitutive rule can equally be described by a set of regulative norms that include the necessary permissions and prohibitions. The existence of a status *Y* simply stands for certain equilibrium relationships among agents "behind it". By shifting Searle's status functions into equilibrium terminology, the unified ontology connects institutional "oughts" to strategic coordination, however, the proper place of "oughts" remains underdeveloped. As they conclude, institutions have a dual function in game-theoretic terms:

- etiological — causal, promoting cooperation
- teleological — evaluative, securing values by imposing social norms (Hindriks & Guala, 2021).

Hindriks and Guala build their account of functions of institutions on the basis of Wright's analysis of biological functions (Wright, 1973), which might be summarized in two main conditions. The first is that the function *F* of an entity *A* is the cause of the existence of *A*. The second condition is that *F* is a consequence of the existence of *A*, which means that *F* is non-redundant to the existence of *A*. The same logic, as the authors argue, applies to institutions. Promotion of cooperation to solve coordination problems is presumably the cause of the persistence of institutions. And securing the normativity of institutions is their purpose.

1 Guala argued elsewhere that Lewisian conventions acquire normativity (or deontic force) in *repeated play*, so that history of play becomes a focal point for emergence of normativity (Guala & Mittone, 2010). This view is *dynamic*, whereas the view of payoff-modifiers as representations of the "built-in" normativity presuppose a static view of institutions. This, along with a couple of other inconsistencies, complicates the explanation of normativity of social institutions with RiE.

Criticisms of Guala's unified social ontology

While innovative, Guala and Hindriks's account has drawn critiques as reducing constitutive rules to equilibria overlooks important aspects of social reality. Key criticisms include the following:

- **Neglect of material and historical aspects**. Rabinowicz praises Guala's broad integration but raises concerns about the equilibrium focus. He points out that treating institutions as "rules-in-equilibrium" can misrepresent their ontology by ignoring material substrates and history (Rabinowicz, 2018). Many institutions like universities, currency systems, markets and traditions involve *concrete people, practices and goods*, not just abstract strategies. Material base of the games played like students and classrooms in a university, as Rabinowicz claims, is essential part of institutions and cannot be abstracted from due to disregarding "basic ontology" of the modeled phenomena. This criticism fails to convince me, as I find the notion of the "basic ontology" misleading due to implying some "folk ontology" and subscription to conceptual analysis which favors intuition of the social world as a valid starting point for ontological scrutiny. Models abstract away inessential parts of systems in question. Although I agree that there are always concrete people and practices with their material bases, including misleading "basic ontology" is inessential for explanation of the structure of social ontology, emergence and stability of its core parts. Another Rabinowicz's criticism is that the set of possible CE is larger than that of NE, thus complicating the problem of equilibrium selection instead of making it easier. It is so for there is infinity of possible probability distributions which can create many CE arrangements. However, seen *dynamically*, as Skyrms (2014) does, for example, correlation of strategies arises in the course of evolution, thus solving the problem of equilibrium selection with symmetry-breaking by chance and stochasticity.
- **Irreducibility of normativity**. Roversi (2021) defends the traditional importance of constitutive rules. He finds the unified account too reductionist: treating constitutive rules as summaries of regulative incentives "strips away" their normative essence (Roversi, 2021). From his perspective, this view fails to explain why individuals feel bound by institutional norms beyond calculative incentives. Roversi insists that one cannot capture the meaning of a status (like professor) by conditional rules alone as they imply a more complex notion of normativity. Contra Roversi, I believe that Guala and Hindriks do not want to *reduce* normativity to instrumental rationality. On the opposite Hindriks (2019)

elaborates on the definition of social institutions as norm-governed social practices and contends that modeling social norms as sanctions with costs that agents incur for violating norms, is insufficient for its perception by agents as legitimate, According to Hindriks, instrumental account fails to capture the motivation by the norm itself and not by the costs of its violation. He claims that it is *normative expectations* and especially *normative beliefs* that complement sanctions as a source for norm existence and perception as legitimate. Social norm governs a practice if its participants are motivated to follow its rule to a noteworthy extent. This view partially departs from RiE and Hindriks even refines RiE as "rules *and* equilibria", thus disentangling normativity from equilibria.

- **Overestimation of coordination**. Hedoin (2016, 2021) has criticized RiE on the grounds of misrepresentation of institutional stability as overly based on coordination. While Guala's game-theoretic theory offers an elegant and parsimonious model of institutions as coordination equilibria, it is ultimately too thin to capture the full social ontology of institutions. Hédoïn's constructive thesis calls for a broader, normatively rich, and historically grounded framework that recognizes institutions as constitutive of agents' preferences and identities, embedded in social and cultural contexts, and endowed with normative powers that generate genuine obligations. This richer approach aims to explain not just how institutions solve coordination problems but how they shape social life at a fundamental level. The main novel idea is to make exogenous preference endogenous, meaning that equilibria do not reflect agents preferences, but *constitute* them. For example, the institution of marriage changes how individuals value relationships, not just how they coordinate behavior. This dimension is missing in Guala's payoff-matrix approach. While this is true, Hedoin's argument weakens in the light of the dynamic view of institutions: both coordination problems and values might have evolved as adaptations to each other, and as Ross (2012) noted, it is hard to say what was first. It means that Guala's theory can accommodate the transformative view of institutions by integrating values into game matrices and saying that institutions change the rules of the game as Vlerick (2016) does. In addition, Hindriks & Guala (2021) argue that one of the main functions of institutions is to preserve values by the means of social norms, which means that RiE at least presupposes integration of values.

Although the aforementioned critiques highlight important drawback of RiE, most of them do not go deep enough to spot a problem in

the very structure of RiE which affects its overall explanatory capacity. As I will argue in the next section, there are intrinsic problems with the core notions of "correlation" and "representation" within RiE which make the introduction of rules and entire explanation based on the comparison of game-theoretic models in human and animal cases problematic.

Problems with "correlation" and "representation" in RiE

What is the justification for connecting rules and equilibria accounts of social institutions? According to Guala (2016), institutions as rules-in-equilibria are normatively-driven behavioral regularities comprising CE. "Rules" here are the recipes guiding and prescribing certain behavior and are used by the agents themselves, and "equilibria" are *objective stable states* of the strategic interaction between agents. Other scholars pinpoint normative and self-sustaining nature of institutions as "humanly devised constraints that shape human interactions" (North, 1990), "norm-governed social practices" (Tuomela, 2013) and "self-sustaining salient behavioral patterns" (Aoki, 2007). Institutions combine "subjective" and "objective" components: they are driven by social norms, that might vary from one population to another, and, at the same time, constrain possible actions and sustain itself.

The rule-based account conceives of social institutions as rules guiding and constraining behavior in social interaction or "humanly devised constraints" of social interactions (North, 1990). In sociology, the tradition of treating institutions as rules dates back to such classical figures as Weber (1924) and Parsons (2015), and it continues to thrive today. The equilibrium-based account sees institutions as behavioral regularities and, most importantly, solutions to coordination problems. The constitutive rules account sees institutions as systems assigning statuses and functions to physical entities like we saw earlier in Searle (1995).

Guala & Hindriks (2015) argue for the reconciliation of rules and equilibria from *explanatory insufficiency* of rules and equilibria as separate explanations of the *stability of social institutions*. The focus on stability of norm-governed mutually beneficial recurrent behavioral patterns recasts the problem of the "unification of ontology" in terms of causal efficacy of the factors influencing stability of norm-governed practices, although Guala and Hindriks do not put this explicitly.

According to Hindriks & Guala (2015), the rule-based account is *insufficient*, for it cannot explain why some rules are followed and others are not. For example, they recall a French law from 1799 which forbid women to legally wear trousers and "dress like men" which was officially revoked only in 2013. The failure of this "rule" (as French women did

wear the trousers despite this law) shows the fragility of the rules-only approach to social institutions.

To address this issue, Hindriks and Guala use an equilibrium account to show the *strategic character of rule-following. The authors compare the two paradigmatic games from game theory, prisoner's dilemma and stag hunt. Although mutual defection in the prisoner's dilemma is a NE, it is not a social institution, for it is not self-sustaining due to independence of players' strategies. In contrast, the mutual decision to hunt a stag instead of a hare, which are also both NE, is an institution, for it requires correlation of players' strategies to achieve a bigger joint payoff. The latter means that the strategy is salient and beneficial for players, which seemingly explains why some rules are followed and others not.

	C	D
C	-1, -1	-3,0
D	0, -3	-2, -2

	S	H
S	4,4	1,3
H	3,1	2,2

Prisoner's dilemma (left) and Stag hunt (right)

However, the notion of players' correlated strategies as an explanation of the stability of institutions is insufficient, as well, as the authors point out, for it is too permissive. They provide an example of non-human animals solving coordination problems without institutions. For example, male baboons, lions, swallowtails and other species exhibit a recurring behavioral pattern that can be described in terms of CE. Males patrol an area to mate with females and have ritual fights with intruders if encountered. The evolved pair of players' strategies minimizes possible damage to both parties and lets the incumbent occupy territory and mate (Maynard Smith, 1982). The authors use Maynard Smith's exposition of animal territorial behavior represented as a "Hawk-Dove-Bourgeois" (or HDB) game to provide a simplified example of a prototypical social institution of property:

	H	D	B
H	-1	2	0,5
D	0	1	0,5
B	-0,5	1,5	1,0

Hawk-Dove-Bourgeois game

Presented with a terrestrial resource, a "Hawk" player fights over it, a "Dove" retreats and "Bourgeois" uses a conditional strategy "fight if own and retreat if do not own". In this game, Guala and Hindriks see the "bourgeois" strategy "fight if own" as a correlated one in the sense of Aumann/Vanderschraaf, meaning that the players coordinate their actions by conditioning them on an external signal. As they say, it is a "simple pre-emption device: whoever occupied the land first has the right to use it" (Hindriks & Guala, 2015, p. 465). In this case, the temporal order of occupation is used as a correlation device. Overall, this correlation fulfills the necessary condition of being an institution.

Guala & Hindriks (2015) illustrate the applicability of HDB to humans with a game where two tribes, spatially separated by a dry river, graze their cattle. The dry river serves as a "focal point"—a salient feature of the environment that the members of both tribes have been aware of (Schelling, 1980). It also serves as a correlation device, for it is a source of a public signal that coordinated actions of different tribes without their explicit agreement. Thus, the shepherds of both tribes have three possible strategies: "Graze", "Not graze" and "Graze if North / South of the river" according to the history of their territorial occupation. The members of one tribe might be killed by the members of another if grazing their cattle on another side of the dry river which the other tribe possesses. The stable set of strategies is grazing if on the own side.

	G	NG	GIS
G	-1	2	0,5
NG	0	1	0,5
GIN	- 0,5	1,5	1,0

Grazing game: the player strategies are Graze, Not Graze and Graze if North / Graze if South

The payoff structure of the game is identical for animal and human cases, and we cannot discriminate between them solely on this basis. This invites the insufficiency of CE employed by Guala and Hindriks. However, although they mention adding payoff modifiers to the game structure to accommodate inherent normativity of conventions (Hindriks, 2019), they do not explain how animal conventions evolve into human payoff-modified ones[1]. So the question is twofold:

1 As Guala notes, coordination in repeated games is achieved via modifying payoffs with parameters to increase the costs of free-riding and violation

- how payoff-modified game is connected with CE?
- where does normativity reside in the Dinka/Nuer game?

These questions are hard to tackle up-front, so we will start with easier ones. As the structure of the HDB game is identical for animals and humans, and this is why equilibrium-only account of social ontology necessitates rules or a theory of normativity in general, what then distinguishes "animal conventions" like butterflies' territorial contest resolution from human social institutions like private property?

Apart from the obvious notions of "normativity" and seen purely from a formal point of view, Guala and Hindriks argue that animal and human conventions differ in the *scope of actionable signals*. Building on the work of Sterelny (2003), they contend that animals may only respond to a limited genetically "wired" set of stimuli, whereas humans, with their ability to use representations and symbols to condition behavior, can *decouple response from stimulus and condition behavior on many representations* which they equate with following rules. For example, butterflies cannot coordinate on anything but who occupied the sunspot first and unable to create new equilibria. Humans, however, can go beyond this by choosing arbitrary correlation devices and expanding the number of equilibria (Guala & Hindriks, 2015).

To summarize Guala's and Hindriks's argument:

1. social institutions are backed by regulative rules instead of genuinely constitutive ones
2. regulative rules are agents' conditional strategies comprising a CE
3. constitutive rules are linguistically transformed regulative rules with added theoretical term that represents a certain equilibrium or a set thereof
4. at the same time, many animal species exhibit conditional behavioral patterns describable as CE
5. despite the similarity of the matrix representation, the cases of "animal conventions" and human social institutions differ in scope of actionable signals, as humans can invent and follow new rules with the help of *representation of environment*, whereas animals are bound to genetically inherited sets of behavioral responses

---

which by itself creates salience in repeated games (personal communication). However, it leaves questions of where do parameters modifying payoffs come from and how correlation of strategies and payoff modification are connected. It seems that normativity is "embedded" in RiE by default without explaining it, but it might be explained if seen dynamically.

6. the arbitrariness of rules that humans can invent and follow is grounded in and ontologically depends on shared representations of a given community.

Put differently, the difference in scope of actionable signals between animals and humans is due to humans having social epistemology with normative beliefs that grounds social ontology.

The "scope of actionable signals" in more detail

What, however, is "increasing number of equilibria" and how exactly the "scope of actionable signals" is different in animals and humans? If both animal and human cases can be seen as CE, it presupposes a probability distribution over possible strategies which "suggests" players to play a certain strategy. The number of strategies is equal in both cases, as well, so "increasing number of equilibria" is not literal and must reside, then, somewhere else. "Scope of actionable signals" means that humans can attend to and correlate strategies on a wider range of correlation devices. Formally, the representation of a probability distribution over strategies characteristic of CE is identical and hence does not grasp the difference between animal and human HDB.

I suggest that what is increasing between animal and human HDB is the *degree of arbitrariness of a correlation device*: to what extent it could have been otherwise? We could quantify it by appealing to *how many features of environment can be salient* for an organism to coordinate upon? In Lewis's language, it is "how many states of the world agents are able to recognize?". Butterflies, for instance, coordinate on the order of precedence among other things, but their set of recognizable salient features is much smaller than humans' (which is not quite surprising, to be honest).

If we view a set of *recognizable world states* for a butterfly schematically, we might get:

Ω = {Spot is occupied,
Spot is free,
Opponent is bigger,
Opponent is smaller}

This set of recognizable world states can be generalized to a set of *salient environmental features* and hence of possible correlation devices *D* which agents can coordinate upon to arrive at CE:

D1 = {precedence, opponent's size}

What *D* represents is the possible sources of signals given by CE to players. If precedence is salient, then CE will consist of probability distributions over strategies "Hawk, Dove, Hawk if you occupied the spot first, Dove if the spot is already occupied". This essentially models ownership as in "owner/intruder" role assignation in Maynard Smith's HDB (Maynard Smith, 1982). If opponent's size is more salient, the CE would contain "Hawk if bigger" and "Dove is smaller" strategies.

However, it is only applicable to one particular case of a territorial contest and not the entire behavioral repertoire of butterflies. This suggests that in other games other sets of possible relevant correlation devices might hold. And the number and nature of these correlation devices is ultimately *bound to the level of cognitive sophistication of an agent*. This is why, I suppose, Guala appeals to Sterelny by saying that "more complex creatures can decouple stimulus and response with representation of environment". This is what I mean by the level of cognitive sophistication affecting the amount of potentially salient environmental features.

To illustrate it, we can classify agents by the number of games they are capable of playing and how big is their set of correlation devices (which equals to how many environmental features they can attend to). For example, for the sake of simplicity, let us consider butterflies as only capable of playing Hawk-Dove-Bourgeois and not "Prisoner's dilemma" or "Stag hunt". Humans can play all sorts of games and the amount of correlation devices in each can be bigger than two. Consider Guala's "Grazing game":

D2 = {precedence,
position to the Sobat Riverbed,
shared knowledge of each others' property,
...}

Both the quantity and the quality of correlation devices differ. What Guala specifically mentions is that humans can condition strategy not on a single cue, but on multiple ones simultaneously, which he equates with following rules. What this practically means is that we, I suggest, can *integrate* many salient cues like "precedence", "shared knowledge of others' borderlines" and others to come up with a strategic response. However, this is a toy example, since it does not reflect a real question which Herrmann & Skyrms (2021) emphasize: "what is the set of asymmetries over which *selection is operating*" and not just all logically possible asymmetries? It reflects the ontic/epistemic divide, since we sketched the sets of asymmetries *recognized* by agents as opposed to those on which *selection operates*.

If we abstract away from a particular case of riverbed-based borderlines, the amount of things which could serve as a correlation device for

coordinating on borders and property would grow, for there is *functional independence* of human correlation devices from a set of environmental features which could possibly be salient. We can *use* different things as correlation device in functionally similar situations like deciding who goes first when pedestrians or cars clash into each other. There is a possibility, however, that there is an exhaustive set of functionally possible correlation devices which can be deployed across contexts varying from more mechanistic like "precedence" and "spatial position" to more socially constructed categories like "age" (since its perception and attitude towards can vary culture-wise), "gender", "role" and others[1].

Two important features differentiate human correlation devices from animal ones here:

- *Functional abstraction*: we can use contextually-accessible environmental cues as functionally similar correlation device (shells, necklaces, coins and cigarettes can be money in different contexts);
- *Beliefs as cues*: we can use beliefs as "environmental cues". For instance, players in the "Grazing game" base their strategies not only on physical markers, but on shared belief about existence of another tribe and consequences of trespassing, although it is not evident from the very formulation of strategies like "In North to the river, graze".

Human institutions are *multiply realizable* in the sense of being functionally abstracted from particular instantiations. Applying this insight to our example of the set of correlation devices D_2 for "Grazing game", we see that the set would consist of different things from context to context as different environmental cues can be relevant for establishing territorial borders, just like money.

Thus, *environmental asymmetries appearing to organisms as salient comprise a set of possible correlation devices.* The greater the number of recognizable asymmetries, be it physical or epistemic, the greater the amount of possible correlation devices. So, what Guala most probably means by "scope of actionable signals" is essentially a set of recognizable salient environmental asymmetries.

Conflation of agency models in RiE

Guala and Hindriks argue that an *adequate* theory of institutions must have three explanatory components (Guala & Hindriks, 2015, p. 469):

1 O'Connor (2019) has studied the emergence of unfairness and gendered labor division in this manner, although she has been criticized for using social categories themselves as correlation devices and not a thing to be explained in the first place (Saunders, 2022).

- coordination
- correlation
- representation,

where "adequateness" is related to the *stability of social institutions* as the main explanatory target.

According to Guala and Hindriks, not all conventions are social institutions, but only those that require solving coordination problems in Lewisian sense of achieving mutually beneficial outcome with *salient* strategies. Salience comes from the normativity of strategies as extra-costs (personal communication). At the same time, salience can be based on a public signal like a dried riverbed of the Sobat River as in the example of Dinka and Nuer tribes. Normativity might, then, arise in RiE from the more "low-level" features of environment over the course of repeated play akin to Ullmann-Margalit's (1977) normative thresholds.

Institutions coordinate behavior by *correlating agents' strategies in coordination games* as in Aumann/Vanderschraaf/Gintis notion of CE, and humans are able to *devise many strategies* and hence equilibria given the same correlation device, or signal. As Guala and Hindriks stipulate, independently chosen strategies like NE in the Battle of the Sexes are not social institutions as they are neither mutually beneficial, nor self-reinforcing. Correlation of strategies shows:

- that agents can coordinate upon any coordination device
- that coordination is mutually beneficial.

As might have been noticed, there is a subtle tension within the game-theoretic part of Guala's and Hindriks's theory. While discussing HDB, they say that animals play CE and at the same time credit Maynard Smith (1982) as for coming up with the very conceptualization of the HDB game. An important remark, though, is that Maynard-Smith used HDB to introduce the concept of ESS as we already know and did not use the notion of CE. An intuitive question is how do CE and ESS hang together in the same game-theoretic description of an animal convention, especially taking into account the conditional nature of CE which animals supposedly lack. As we remember, CE is given by:

$$\sum_{s_{-i} \in S_{-i}} p(s_i, s_{-i}) u_i(s_i, s_{-i}) \geq \sum_{s_{-i} \in S_{-i}} p(s_i, s_{-i}) u_i(s_i', s_{-i}), \quad \forall s_i, s_i' \in S_i$$

$$u(s^*, s^*) > u(s', s^*) \quad or \quad [u(s^*, s^*) = u(s', s^*) \quad and \quad u(s^*, s') > u(s', s')].$$

And ESS is given by:

which means that, although they both are Nash equilibria, they have diverging assumptions. For RiE, it means that:

- either "animal conventions" are mathematically different from human social institutions, for they represent ESS and not CE, and

hence Guala's argument from insufficiency and the very reason for introduction of rules seems to fail (for one misinterprets ESS for CE with its rationality assumptions);
- or "animal conventions" are themselves correlated, and there comes the need of showing *in what respect these correlations are different*. Only by showing it we can disambiguate the claims of RiE about CE as its backbone.

Guala and Hindriks use the notion of *"bourgeois equilibrium"* (BE) introduced in the context of evolutionary game theory by Maynard Smith (1982). They see the HDB game of animal territorial ownership a prototypical "animal convention". According to the authors, the so-called BE is essentially a CE, however Maynard Smith used BE to define ESS. It creates tension, for CE and ESS are mathematically distinct: the former is "too loose" and the latter is "too strict" in terms of the stability conditions (and RiE strives to explain the stability of social institutions), and it is unclear how they can be combined. Hence, the issue consists of clarifying the status of BE: whether it is a CE, an ESS or something else. It is due to being at the core of Guala's argument for institutions as correlation of strategies rooted in evolution.

Guala's "Grazing game" presupposes that both tribes know that the other tribe exist over the river and they know the consequences of crossing the river as they are embedded into the game structure. This "consequentialist" position suggests that agents correlate their strategies not on interactional asymmetries or simple environmental cues, but on higher-order representations. And in spite of Guala's idea that agents use simple heuristics and conditional rules like "Graze if North", it seems like such heuristics are the result of previous more cognitively demanding reasoning and not a first-time optimization. The very payoff structure and the notion of both tribes' knowledge of it presupposes at least some proto-normativity, as shared knowledge of low payoff for a certain strategy signals to agents not to use it.

The ambiguity over what is actually a "bourgeois equilibrium" calls for an inquiry into its origins in Maynard Smith's evolutionary game theory and subsequent comparison with Guala's account.

Ontic and epistemic correlation of strategies

Maynard Smith (1982) famously has introduced the notion of ESS into game theory. A 'strategy' is a behavioral *phenotype*, a specification of what an individual will do in any situation. An ESS is a strategy that, if adopted by all members of a population, prevents the invasion of any mutant strategy by natural selection. The concept originated in the context of animal behavior, but can be applied to any phenotypic variation.

Maynard Smith proposed a model of a 'Hawk-Dove' game that represents a mis-coordination game between two agents. In a competition for some resource, 'Hawk' fights for it and 'Dove' displays and retreats if threatened.

	Hawk	Dove
Hawk	$\frac{V-C}{2}$ $\frac{V-C}{2}$	V, 0
Dove	0, V	$\frac{V}{2}$ $\frac{V}{2}$

A 'Hawk-Dove' game. The payoffs are determined by the value of the resource (*V*) and the cost of fighting (*C*). Value *V* increases the Darwinian fitness of an individual if they obtain the resource, and cost *C* reduces it if injured in a fight over the resource. Not gaining *V*, however, does not mean zero fitness.

This model assumes:

- an *infinite population* with asexual reproduction
- symmetric contests between two opponents
- random matching
- a finite set of strategies.

Defining the stability criteria for the strategies, he proposed that If a strategy *I* is stable against *J*, it must satisfy the "standard conditions" from Smith & Price (1973): the fitness of typical members adopting *I* must be greater than any mutant *J*, such that:

- either $E(I–I) > E(I–J)$ *or* $E(I–I) = E(I–J)$
- and $E(I–J) > E(J–J)$.

According to these conditions, *D* cannot be an ESS, for $E(D, D) < E(H, D)$, and *H* is an ESS if costs of injury are less than potential gain from the resource, $V > C$. If $V < C$, neither *H* nor *D* is an ESS. To proceed, Maynard Smith considers the behavior of individuals who can play either strategy with a certain probability, which they pass on to their offspring. This strategy takes the form 'play *H* with probability *P*, and *D* with probability (1–*P*)'.

A mixed strategy *I*, which randomly chooses an action from a set of possible actions, may be an ESS if the expected payoffs of the strategies composing it are equal. This follows from a theorem by Bishop & Cannings (1978): if a mixed ESS includes the pure strategies *A, B, C...* with non-zero probability, then

$$E(A, I) = E(B, I) = E(C, I) = \ldots = E(I, I)$$

Intuitively, this means that if $E(A, I) > E(B, I)$, adopting *A* more often and *B* less often would be more advantageous than following strategy *I*, making it not an ESS.

However, *I* can be an ESS if probability of its adoption is *P = V/C*. In contests where the cost of injury is greater than the rewards of victory, *V < C*, mixed strategies with *P = V/C* are evolutionarily stable.

What is important, a game with two pure strategies always has an ESS, and games with three or more strategies may not have one. As we remember, both "Hawk-Dove" and "Grazing" games have three strategies.

Maynard Smith (1982) introduces the distinction between *symmetric and asymmetric games*. He illustrates them with animal contests. An asymmetric contest is one where participants have different roles, allowing them to use different strategies. Roles must be identifiable and can be based on gender, ownership, or intruder status. Circumstances which determine an individual's role are assumed to be independent of their genetic strategy. A contest with no role differentiation is 'symmetric'. Maynard Smith (1982) characterizes them as follows:

Contests are between two individuals of distinct roles (e.g., owner/intruder, larger/smaller, older/younger);

Both individuals know their role;

Both have the same strategies available (e.g., escalate, retaliate, display);

Role may influence chances of winning or value of victory.

The Hawk-Dove game is symmetrical—both players have the same choice of strategies and payoffs. However, most contests are asymmetric, with differences in size, strength, gender, age, or ownership influencing strategy choice and/or altering payoffs or success in escalation. Even when the asymmetry does not change payoffs or escalation outcomes, it may still determine the players' actions.

	Hawk	Dove	Bourgeois
Hawk	-1	2	0,5
Dove	0	1	0,5
Bourgeois	-0,5	1,5	1

In this example, the Hawk-Dove game is extended to include a third strategy, *B* (or Bourgeois), which is defined as 'if owner, play Hawk; if intruder, play Dove'. This strategy is ESS and the only ESS of this game. It is assumed that each strategy type is owner and intruder equally frequently. Hence, even when ownership does not alter payoffs or success in fighting, an asymmetry of ownership can be used as a conventional one to settle the contest. It is a correlation of strategies.

Here, the *B* player chooses *H* and *D* with equal frequency, acting as an owner on half the occasions and an intruder on the other half. And when two *B*'s compete, if one chooses *H*, the other chooses *D*. If *V > C*, the ESS is *H* as it is worth risking injury to gain the resource; if *V < C*, the ESS is

B as ownership settles the contest without escalation. It means that in both 'Hawk-Dove' and 'Grazing' games $V < C$.[1].

Crucially, this assumes that *the probability of an individual occupying a role is independent of their strategy*. This holds true even for strategy *B*, wherein the individual's role is correlated with their chosen action (Hawk or Dove). The assumption is that the strategy *B* itself is unrelated to role. In other words, If an agent is indeed an 'owner', it does not entail that she always plays a certain 'owner' strategy like 'Hawk' or 'Bourgeois'. However, according to Gintis (2007b), empirical findings corroborate the existence of the 'endowment' effect, when owners value a resource more than intruders, thus making them fight harder for it. It presupposes a certain degree of correlation between role and strategy, as opposed to Maynard Smith's purely theoretical model.

Smith & Parker (1976) used the term 'uncorrelated asymmetry' to refer to payoff-irrelevant differences in an otherwise symmetric game. In asymmetric contests the value of the resource, or chance of victory, is not the same for both owner and intruder. The payoffs to owners and intruders are often not equal, so the territory may be more valuable to an owner who has already familiarized themselves with food, refuge, and other. Ownership may even offer advantages in escalated contests. Inequality of payoffs is possible due to size or age asymmetry. Even if there is no inequality, an asymmetry can still settle contests. Thus, "Grazing" game as presented by Guala, does not require a correlated device and may be solved by uncorrelated asymmetry alone, as both players recognize the asymmetry of ownership and the value of territorial gains is less than the costs of potential injury, $V < C$, for they might value their own territory more than potential one.

It is interesting that Maynard Smith (1982) considers the 'social contract' game as one which humans can play but animals cannot. This game involves a group of individuals agreeing on a behavioral regularity and punishing any member who deviates from it, thus implying "full-fledged" normativity. However, the act of punishing carries a cost, so in order to maintain stability, refusal to participate in enforcement must

1 However, it is not clear whether human players such as grazers have genuine fitness rather than utility value function. As Sterelny (2012a) suggests, there has been an evolutionary shift from fitness to utility correlated with the demographic explosion in the Pleistocene and subsequent significant decline in individual-level heritability of cultural traits, for offspring did not more resemble their parents informationally and ideologically due to the abundance of cultural information sources. He argues for model pluralism, although we are inclined to view humans agents as rather utility maximizers that fitness ones. A critical examination of the very metaphor of an individual as a maximizing agents is aptly expressed in Paternotte (2020).

be considered a breach and punished as well. To ensure enforcement, a subgroup may be rewarded for carrying it out.

Essentially, the 'social contract' game differs from any other coordination or cooperation game in added normativity which is implicit in the payoff structure. Such normativity is characteristic of social institutions as defined in rules-in-equilibria theory.

Guala and Hindriks put forward that coordination in social institutions, and in HDB as an exemplar case of property, is due to correlation of strategies. But what is "correlation of strategies" in the first place?

Correlation of strategies is a stable state of strategic interaction. It is represented by CE. It allows players to make decisions based on an external signal from a joint probability distribution over possible strategies. As per Aumann (1987), CE is based on beliefs about how the other players will act, which can increase the efficiency of their own strategies. The concept of CE has been used to explain various phenomena in strategic decision making, including how people form coalitions, how firms cooperate and compete, and how players interact in team games.

The key feature and difference of CE is randomization. As Aumann (1987) points out, correlation is more general than mixing of strategies, for the latter can be formally seen as the former by considering the product probability space $\Gamma^1 \times ... \times \Gamma^n$, where Γ^i is the set of outcomes corresponding to player *i*'s mixed strategy. Players make correlated, or nonindependent, choices when they observe the same random variable.

In the HDB game, one of its important features is that if $V < C$, the *B* strategy helps to settle contests *conventionally*. Maynard Smith does not emphasize the notion of convention, but he probably meant by it is that there is shared 'understanding' of the situation expressed as a joint probability distribution over possible strategies between the players that helps to choose a strategy with a signal.

One intrinsic problem with *BB* as CE, however, is the source of randomization. Some scholars appeal to Nature as to such source and correlation device in the animal case, thus eliminating the tension between the requirement of randomization and symmetry of ESS (Cripps, 1991; Metzger, 2018; Skyrms, 2014). In particular, Gintis defines CE as an NE of a game *G* augmented by the *initial move by Nature* that observes a random variable *Y* on a probability space (Γ, p) and issues directives $f_i(\gamma) \in S$ to each player *i*, such that choosing the directive is the best response given agents having a common prior *P* and assuming other players are also following Nature's directives (Gintis, 2009, pp. 135–136). The crucial assumption, though, is that *the game is epistemic and has common priors*. This implies that all agents agree on the probability distributions over their actions and have *joint randomized probability*. Assuming a strict CE—that each agent *i* has a single best response $S_i(\omega)$ in every state ω —each player's move is known to the others and rationality dictates that each

must play a best response to the actions of the others. If we plug here our earlier example of sets of correlation devices, we can represent γ as a composite cue vector $\gamma = \{opponent'ssize, precedence...\}$ which encode all the relevant information about the world state.

In their theory, Guala and Hindriks endorse the epistemic interpretation of the "Hawk-Dove". It means that "animal conventions" are themselves correlated, but it is still confusing in regard to the diverging assumptions of evolutionary and epistemic interpretations of the game. In other words, how could animals correlate their strategies non-epistemically while still having ESS?

According to Skyrms, the conditional nature of the *B* strategy is genuinely correlative: the *BB* strategy profile is a CE *spontaneously arising* from symmetry-breaking that happens when individuals randomize the choice of their strategies and do not know whether they are "Hawks" or "Doves" (2014, p. 78).

This is where Skyrms's notion of "correlated convention" shines. To reiterate, correlation in evolutionary game is achieved with the symmetry broken with interactional dependencies between agents (spatial or social structure, signaling, etc). This is possible due to a frequentist interpretation of probability in CE without beliefs instead of subjective one.

A random mutation or learning can create a CE in a Hawk-Dove game. A slight increase in resource value or fighting ability will strongly favor the BE over other strategies. What was "uncorrelated asymmetry" for Maynard Smith, became "correlated asymmetry" for Skyrms.

On Skyrms's account, *BB* is CE *and* ESS at the same time without a conflict (personal communication). More precisely, *BB* is an *adaptive-ratifiable* strategy (generalized ESS), for it takes into account conditional strategies emerging from endogenous structure of interactions and resulting in correlation which affects interaction frequencies and stabilizes the strategy over the population. This strategy does not assume *any* rationality in agents, let alone a Bayesian or a boundedly rational one implied by Guala.

For RiE, the notion of "correlated convention" means that an example of animal HDB can indeed be described as CE and ESS at the same time under the crucial condition of the *non-epistemic nature of the game*, whereas Guala's agents are strictly epistemic (personal communication).

The crucial difference is that non-epistemic agents correlate on the *objective features of environment*, while epistemic agents *correlate on beliefs*, especially in the case of human HDB with added normativity, where social norms are the source of salience like Guala and Gintis put forward (Gintis, 2009; Guala & Hindriks, 2015).

We might introduce the distinction of *ontic* and *epistemic* correlation. what is subtle yet crucial here and what is overlooked in RiE, is that

these types of correlation have distinct real-world references. While both animal and human conventions are CE formally and can be described with a single model, both their "inner workings" and references "in the world" are different. this renders the entire Guala's argument of "scope of actionable signals" problematic, for it is not evident how to compare games with ontic and epistemic cues. To remind, this argument is at the backbone of RiE as it justifies the introduction of rules into equilibrium theories of social institutions.

According to Guala and Hindriks, stability of social institutions cannot be based only on CE as it is equally applicable to animals and humans. However, these are two distinct CEs — frequentist (ontic) and Bayesian (epistemic) ones. While rules or any notion of normativity is missing in the ontic correlation by definition (as it cannot be a feature of environment), rules as salient norms can and sometimes are CE themselves. For example, Gintis (2009) views social norms as "choreographers" and CE as they are essentially "second-order" conditional strategies. We can imagine a pair of strategies like "Others expect me to stop if red" and "Others expect me to drive if green" which comprise a CE.

It means that Guala's and Hindriks's argument for insufficiency of CE without rules for human cases is unsubstantiated, for CE as norms already includes both Bayesian rationality *and* normativity. Consequently, explanation of stability of institutions might not require separate rules as norms, as for epistemic games with conditional agents, norms can be CE themselves.

While I am not denying the effectiveness of having both rules and equilibria, I argue that the "insufficiency" argument for introducing norms as separate entities from CEs seems to not go well. Norms are smuggled into RiE almost *ex nihilo* and remain unexplained why we need them for the explanation of stability of institutions, while having them as the central causal factor of such stability according to the insufficiency of CE argument.

Instead, what remains to be explained by RiE is the gap between ontic and epistemic correlation: how stability of human institutions evolves from stability of animal conventions?

Guala and Hindriks mention *representation* as the third component of an adequate theory of social institutions along with coordination and correlation of strategies. Their argument goes that representation helps agents to condition their strategies not on immediate features of environment like in ontic correlation, but on many representations which Guala and Hindriks, importantly, equate with following rules (which is essentially epistemic correlation). So, representation for them is a mechanism for bridging animal and human conventions and explaining emergence of normativity and hence the stability of institutions. However, it is not evident how the two are connected.



A subtle thing about "representation" is that agents represent equilibria with rules, so rules might not even have any causal efficacy as they can misrepresent an actual equilibrium (Guala, personal communication). Rules would have causal efficacy if they contributed to the maintenance of an equilibrium, but Guala claims that these can be misaligned while still have effective equilibria. It presupposes that rules in rules-in-equilibria might be eliminable.

Although I do not advocate for eliminating rules, I believe they are not causally efficacious in RiE, since they only serve as a placeholder to explain why agents exhibit rule-like social behavior (because they internalize objective behavioral patterns and represent them as recipes for action). However, there might be another, causal sense in which representation in RiE could work and do real explanatory work.

So far, we have identified the problem of conflated agency models within RiE which influences the cohesiveness of one of its main arguments. Namely, the argument that CEs are insufficient for explanation of stability of human social institutions, for they are applicable to animal coordination problems and hence necessitate rules as separate explanatory entities. Identification of this problem allows us to view normativity as left unexplained in RiE, for animal CEs are stable without rules and human CEs might themselves be salient social norms as rules. If normativity is already included into CEs, how did we go from non-normative CEs to normative ones? A capacity for representation is responsible for transitioning from the former to the latter, which is a big bet for a theory of stability of institutions.

The problem with "representation" in RiE

The notion of representation employed by Guala and Hindriks is ambiguous, which complicates substantiation of normativity as the main causal factor of stability of social institutions.

While justifying the insufficiency of equilibria with applicability of the concept of CE to both humans and animals, Guala and Hindriks use the notion of representation in a broad sense, although appeal to Sterelny (2003), who uses it in a narrower sense of an *advanced cognitive capacity* evolved in hominids. Coordination by correlation alone is insufficient for introducing distinction between animal and human conventions with normativity, and representation is needed as mechanism for it, as Guala and Hindriks argue.

Correlation in RiE (which we now know is epistemic) is possible, as Guala and Hindriks put forward, due to an advanced *cognitive capacity for conditioning behavior on representation of the environment*. As RiE has "rules" and "equilibria" parts, they must be somehow connected. For this reason, rules in RiE are symbolic *representations* of strategies in a game. They not only serve as symbolic markers of the properties of equilibria, but considerably save cognitive effort. As Aoki notes (2007, p. 6):

> "An institution is a self-sustaining, salient pattern of social interaction, as represented by meaningful rules that every agent knows, and incorporated as agents' shared beliefs about the ways the game is to be played".

If agents *represent* objective behavioral patterns as meaningful recipes for action and incorporate these patterns as shared beliefs, it creates an interesting relationship: agents essentially *share representations of the game structure*, but they also *represent it with recipes for action* for themselves. It means there might be two different representations at play: shared and private ones, and Guala and Hindriks seemingly talk about a private one as responsible for normativity of institutions (although it is not evident for sure).

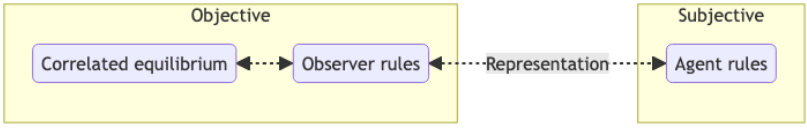
If *rules* are private representations of objective game structure, what are they? As Guala and Hindriks stipulate, *rules essentially are norms*. Guala, following North (1990), notes that social norms make behavior more stable and more predictable. However, as noted by Vlerick (2019), they introduce new behaviors, as well, and they do it by changing game payoffs.

There are two types of rules in RiE:

- *agent-rules* which agents formulate to represent and guide their behavior like recipes
- *observer-rules* which an observer formulates to represent and summarize others' behavior.

Strategies of the game can be described as rules, but a-rules influence behavior and o-rules only describe it (Guala, 2016, p. 54).

Institutions are composed of both subjective and objective components: they are determined by varying social norms as rules and simultaneously restrict certain behaviors and their own perpetuation.



What Guala and Hindriks propose with introduction of representation as a necessary component of a theory of social institutions is es-

sentially a way of connecting objective equilibria and subjective rules. And a core idea is that the latter can misrepresent the former as I have already noted: agents can have false beliefs about objective states.

However, it is not evident how rules *(mis)represent* strategies, although we have suggested that there are public and private representation components in RiE which might not coincide. To clarify, the authors, drawing on Hindriks (2005), bridge their rules-in-equilibria account of institutions with the constitutive rules account of Searle (1995). The latter presents institutions as systems of statuses and functions as the formula "X counts as Y in C". Searle draws a sharp distinction between constitutive and regulative rules, emphasizing the difference in their syntax, for that of the latter is "do X if Y".

The constitutive rules approach argues that our beliefs are essential for the existence of institutions, which involve more than just actions. This applies to objects, persons, and events too. X in "X counts as Y in C" can be replaced by predicates that refer to any ontological category (Guala & Hindriks, 2015, p. 470).

As the authors note, constitutive rules are linguistically transformed regulative rules, aided with a new term to name an institution. Combining these accounts enables researchers to investigate Y terms like "money" used by individuals in everyday life and analyze their internal regulative and strategic character, thus *bridging explicit ontology of social science and implicit ontology of ordinary language*. The main idea of this argument, thus, is that constitutive rules can be developed at will from regulative rules or game-theoretic strategies by introducing institutional terms (Guala & Hindriks, 2015, p. 477).

Regulative rule (if X, do Y)
+ Institutional terms (e.g."money")
= Constitutive rule (X as Y in C)

However, the very notion of "representation" as used in RiE is ambiguous: a-rules as semantic constructs "inside of the heads of agents" seem to "represent" game-theoretic strategies just like the scientific concept of a "gene" represents a physically realizable macro-molecule. In other words, it is not a *mental state*, but a *scientific representation* of an object in the ontological structure of the world with a concept representing it. At the same time, authors mention terms like "stimuli", "behavior" and "representation" in one sentence, implying a narrower cognitive perspective.

From a scientific point of view, representation as a relation makes sense, for it allows investigation of Y-terms like "money" used by agents (Guala, 2016, Chapter 14). However, representation as a cognitive capacity does not have any immediate practical application, especially in so-

ciological data. Hence, we can discern two notions of representation in Guala's and Hindriks' argument:

- representation as relation between a concept and real-world entity
- representation as cognitive capacity for having mental states.

If, according to the authors, representation as a cognitive capacity distinguishes human conventions from animal ones, which is a crucial step in their argument from insufficiency of both rules and equilibria, representation as a relation between the rules and equilibrium might ontologically depend on representation as a cognitive capacity.

According to Guala, agents directly represent the strategies of the game with a-rules (personal communication), which seems to contradict Sterelny's notion of representation used by Guala, as the latter presupposes *representing the immediate features of environment*, and strategies (specifically as strategies and not rules) are abstract representations rather than features of agents' environment.

As the authors base their argument on Sterelny's, the capacity for inventing and following new normative rules depends on response breadth and decoupled representation of environment accessible to humans. However, crucially, there is no explicit conceptual link between representation as a cognitive capacity that grounds rule invention and representation of strategies by a-rules. The former is a feature of agents, and the latter is a feature of a theory describing the agents.

When the authors introduce representation as a final condition for satisfactory theory of institutions along with coordination and correlation, while they intend to mean cognitive capacity (personal communication with Guala), they convey scientific representation. it is because they use it to clarify and justify the relationship between the two parts of the theory: rules and equilibria. Representation here means that agents are capable of representing equilibria and their salient features in symbolic form (Hindriks & Guala, 2015, p. 466). This is possible due to an advanced cognitive capacity for decoupling a stimulus and behavior with the aid of representation of environment. This decoupling allows for conditioning behavior, or strategies, *on many coordination devices*, and the authors take it for humans to be equivalent to "following different rules". Here rules are symbolic representations of the strategies "that ought to be followed in a given game" (Hindriks & Guala, 2015, p. 467).

The argument presupposes that behavior conditioning, and hence strategy selection, occurs already being based on existing rules. To follow a rule, it should already exist in the structure of the game. At the same time, these rules are a-rules, and they already represent existing strategies "that ought to be followed in a given game". It means that behavior is conditioned on existing strategies, and this involves something like a reductive regress: inventing new rules requires not only a

capacity for stimulus-behavior decoupling, but existing equilibria, for here salient features of existing equilibria are used as coordination devices. Although it can be problematic only from a logical point of view and not from a pragmatic and evolutionary one, since realistically we cannot think of a situation with no strategic equilibria (maybe before life?).

In other words, the authors equate representation of salient features of the environment with representation of existing strategies, or behavioral responses, that pre-exist in the current game structure and "ought to be followed". It means that agents directly represent game structure with the aid of a-rules and institutional terms, which does not allow them to form whimsical private (mis)representations.

Decoupled representation is used as a bridge between a-rules and o-rules, but it would mean that stimuli are themselves o-rules of the form "do X if Y". There seems to be a missing link as from an evolutionary point of view, abstract entities like conditional strategies cannot be features of environment to coordinate on. Guala and Hindriks thus put the cart before the horse by saying that "agents represent existing normativity with recipes for action" and do not explain *how representation causally affects normativity* and hence the stability of social institutions.

Cognitive representation, hence, does not do useful work in RiE, for:

1. it is interpreted by the authors as scientific representation relating concepts rather than features of environment
2. it is used to represent already existing normativity in the form of conditional strategies (o-rules).

Would this work without representation as a cognitive capacity? Probably no, for stimulus-behavior decoupling is key for a capacity to invent and follow new rules which distinguishes human social institutions and animal conventions. The introduction of decoupled representation as a cognitive capacity is only due to justifying this difference: although the payoff structure in both HDB games is the same, human agents are able to devise and converge on new equilibria with new coordination devices. If butterflies coordinate only on the precedence of occupying the sun spot, humans are not genetically hardwired for using one and only coordination device for a given situation: we can interpret the same coordination device differently and use various devices for the same function (we called it *functional abstraction* earlier). As an example, many countries have a nod as "yes" and head shake as "no". However, it is the opposite in Bulgaria. A set of signals is the same, but the equilibrium is different. If a foreigner tries to understand a native, the equilibrium might crash. Overall, the argument will not succeed without representation as a cognitive capacity, for there will still be no difference between human social institutions and animal conventions in game-theoretic terms.

And would the argument work without the notion of representation as a relation between rules and equilibria? Probably no, as well, as it is the crux of the argument for unification done by RiE. Representation here logically connects rules and equilibria and helps to further connect it to constitutive rules theory by the notion of institutional rules, or Y-terms, as in "X count as Y in C" formula.

A more interesting question is whether representation as relation is possible without representation as a capacity. Again, probably no, for as there is no structural difference between animal conventions and human social institutions without a human capacity for stimulus-behavior decoupling, there is no added representation of strategies with a-rules by agents. Animals seemingly cannot represent strategies with formulated normative a-rules (otherwise they would also have complex institutions as we do). And if there is no decoupling, there is no "new rules, strategies and correlation devices".

Apart from this, according to the authors, representation is needed to condition the behavior on the features of existing equilibria "that ought to be followed in a given game" to introduce new strategies and equilibria. It means that behavior conditioning, either in Sterelny's sense of salient features of immediate environment or in Guala's and Hindriks' sense of a-rules as representations of salient features of existing equilibria, requires a capacity for a decoupled representation.

Thus, for the argument about social institution as rules-in-equilibria to succeed, Guala and Hindriks need to show how representation as a relation between rules and equilibrium is related to representation as a cognitive capacity. In stricter terms, they need to show how the former is dependent on the latter, for this is the only way to *ground* normativity, which is why they initially introduced "representation". This picture would only make cognitive representations causal and not "rules" from RiE and they will remain heuristic, while actual causation would happen at the level of representations of environment.

For RiE to work, the authors need to clarify the *mechanics of decoupled representation*: how it contributes to the emergence of new strategies to the extent that agents acquire an advanced capacity to represent game structure and salient features of equilibria, which supposedly leads to normativity which stabilizes human CEs.

We wrap this paper with a brief discussion of the notion of scientific explanation and the deeper problem with RiE related to it. RiE employs representational explanation, which might have led to the conflation of agency models and supposedly heuristic (non-causal) use of "rules". Using an ontic (causal mechanistic) explanation might help untangle the problems of RiE and build a pathway to a more naturalistic social ontology.

Representational explanation as source of problems in RiE

What we have identified as a conflation of agency models can be described in more general terms of scientific explanation and different view thereof.

According to a prominent family of views in the philosophy of science, scientific explanations fundamentally involve subsuming a phenomenon or its description under a general representation. This *representational subsumption view* (RSV) has been articulated in various forms: for Hempel (1965), explanations take the form of generalizations expressed in first-order logic; Kitcher (n.d., 1981) conceives them as argument schemas; Bechtel & Abrahamsen (2005) emphasize mental models; Churchland (1992) appeals to prototype vectors[1]; and Machamer et al. (2000) focus on mechanism schemas. Despite differences in the specific representational format, these accounts share the core assumption that the philosophical debate about scientific explanation should center on the nature of these representational forms.

While the RSV plausibly captures important aspects of human understanding and how scientists communicate knowledge, it is precisely this focus on representations that Craver (2014) argues is misplaced as the starting point for a philosophical theory of explanation. Instead, he contends that philosophy must shift attention away from the structure of representations toward the ontic commitments about the real-world structures that genuinely count as explanatory. In other words, explanations are not merely *representations* or models but are *grounded in objective causal mechanisms*: entities and activities organized to produce the phenomena to be explained.

The distinction between ontic and representational accounts of scientific explanation shows a fundamental philosophical divide in understanding how scientific explanations function as they have diverging underlying assumptions. The ontic perspective posits that explanations are not primarily linguistic or cognitive entities but exist in the objective causal structures of the world itself. From this viewpoint, a philosophical theory of scientific explanation must achieve two critical goals:

1 Prototype vectors refer to patterns of neural activation in the brain that represent the central tendencies of categories or concepts. Instead of relying on symbolic or rule-based representations, the brain encodes concepts as points or regions in high-dimensional neural state spaces formed through experience. Scientific explanation, on this view, involves mapping new phenomena onto these prototype patterns via neural computation and pattern recognition, emphasizing a sub-symbolic, biologically grounded account of understanding (Churchland, 1992).

explanatory demarcation and explanatory normativity. The former distinguishes explanation from other scientific achievements like "control, description, measurement, prediction, and taxonomy," while the latter determines "when putative explanations succeed and fail" (Craver, 2014). The proper question to models on ontic account should be "what our models should represent?".

Applying the ontic/representational distinction of explanation to RiE, we can see that the explanatory power of RiE mainly comes from the game-theoretic model of coordination as a *scientific representation* and not ontic structure of the world. Seeing an HDB model as a source of explanation naturally affords for conflation of agency models, for we can distinguish them only considering *what they do* rather than *how they look*.

The ontic account of explanation in RiE which derives explanatory power from the very structure of causal mechanisms would require explicating such mechanisms which are responsible for normativity of human social institutions as its distinctive feature. To attempt it, we will take Sterelny's theory of decoupled representation which Guala built the "scope of actionable signals" argument on at face value and attempt to explain the emergence of normativity in epistemic agents from non-epistemic ones with a epistemic niche construction theory (Seitz, 2020; Sterelny, 2003; Sterelny, 2012b).

Conclusion

In this paper, we have put the notion of social conventions into the context of social ontology and identified the problems within the leading unified framework of "rules-in-equilibria" theory of Guala and Hindriks.

We have looked at the standard notions of social ontology with Searle's theory of constitutive rules and a more general characterization of this mode of reasoning as the "Standard Model of Social Ontology" with its reflexivity, performativity and collective intentionality; we scrutinized the "rules-in-equilibria" theory with its relation to both "Standard Model" and game theory, looked at its criticisms and, most importantly, identified its deeper problems.

We spotted a problem with agency models in HDB game where Guala and Hindriks illegitimately transferred CE from evolutionary context to an epistemic one. To reiterate, animal HDB in Maynard Smith from whom Guala and Hindriks take the notion of HDB, although it can be seen as CE, *operates on phenotypes* or non-epistemic agents and not boundedly rational Bayesian individuals as Guala and Hindriks propose. This conflation generates a problem in their insufficiency argument for introduction of rules as separate explanatory entities: in human HDB, normativity is already there, while in animal HDB rules are just not needed as a correlation device. I argued that this generates

a further problem with introduction of the notion of "representation" in RiE as an explanation of the difference between animal and human conventions, for Guala and Hindriks seem to conflate scientific and cognitive notions of representation which leaves normativity unexplained and the gap between animal and human conventions mysterious. Importantly, I suggested that the "scope of actionable signals" can be quantified as a set of potentially salient environmental features which can be used as correlation devices, and that humans have more functionally abstract correlation devices. I also argued that problems of RiE might be due to the authors' (probably implicit) subscription to a representational subsumption view of scientific explanation, according to which the explanatory power of a model comes from its representational features rather than from its real-world counterpart of entities and causal processes responsible for an outcome.

Overall, reframing the explanatory target of RiE from the stability of social institutions to the transition from the stable animal conventions to stable humans ones for the sake of resolving the epistemological problem within RiE, make the problem harder. It is because we now we need to tackle the problem of emergence of human social epistemology, which is much grander project than "a social ontology theory from a philosophy of science point of view". The dynamic trajectory of social coordination is complex and, as Ross (2012) points out, social coordination is inherently difficult due to the multiplicity of equilibria, and features of social organization tend to evolve and stabilize to make it easier. However, as he says, it does not illuminate anything about the relationship between coordination problems and emergence of normativity:

"Did socialization come first in hominid evolution, giving rise to complex coordination problems that higher intelligence then evolved to solve? Or, in light of the importance of social norms and conventions for selecting focal points in coordination games, did intelligence generate coordination problems to which increasingly complex social structures are adaptations?" (Ross, 2012, p. 484).

What we can do next is to model the possible evolutionary trajectories of cognitively sophisticated and functionally abstracted correlated devices: from non-normative and non-epistemic agents playing CE in HDB to human social institutions. We can addressing two weaker yet explanatorily crucial parts of RiE: relationship between animal and human conventions as CE and emergence of normative human conventions with decoupled representation. But this is for another time.

Список источников / References

Aoki M. (2007) Endogenizing institutions and institutional changes. *Journal of Institutional Economics*, *3*(1), pp. 1–31. https://doi.org/10.1017/S1744137406000531

Archer M. S. (1995) *Realist social theory: The morphogenetic approach*. Cambridge university press.

Ásta (2024) Social Ontology: Where Now? *Journal of Social Ontology*, *10*(3), pp. 7–10. https://doi.org/10.25365/JSO-2024-9028

Aumann R. J. (1987) Correlated Equilibrium as an Expression of Bayesian Rationality. *Econometrica*, *55*(1), 1. https://doi.org/10.2307/1911154

Baker L. R. (2019) Just What is Social Ontology? *Journal of Social Ontology*, *5*(1), pp. 1–12. https://doi.org/10.1515/jso-2019-2001

Bechtel W., & Abrahamsen A. (2005) Explanation: a mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, Mechanisms in Biology*, *36*(2), pp. 421–441. https://doi.org/10.1016/j.shpsc.2005.03.010

Bicchieri C. (2005) *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press. https://doi.org/10.1017/CBO9780511616037

Bickhard M. H. (2008) Social Ontology as Convention. *Topoi*, *27*(1–2), pp. 139–149. https://doi.org/10.1007/s11245-008-9036-1

Birch J. (2017) *The philosophy of social evolution*. Oxford University Press.

Bishop D. T., & Cannings C. (1978) A generalized war of attrition. *Journal of Theoretical Biology*, *70*(1), pp. 85–124. https://doi.org/10.1016/0022-5193(78)90304-1

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), pp. 127–148. https://www.jstor.org/stable/4320174

Churchland P. M. (1992) *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. MIT Press.

Craver C. F. (2014) The Ontic Account of Scientific Explanation. In M. I. Kaiser, O. R. Scholz, D. Plenge, & A. Hüttemann (Eds), *Explanation in the Special Sciences: The Case of Biology and History* (pp. 27–52). Springer Netherlands. https://doi.org/10.1007/978-94-007-7563-3_2

Crawford S. E. S., & Ostrom E. (1995) A Grammar of Institutions. *American Political Science Review*, *89*(3), pp. 582–600. https://doi.org/10.2307/2082975

Cripps M. (1991) Correlated equilibria and evolutionary stability. *Journal of Economic Theory*, *55*(2), pp. 428–434. https://doi.org/10.1016/0022-0531(91)90048-9

Elder-Vass D. (2007) A Method for Social Ontology: Iterating Ontology and Social Research. *Journal of Critical Realism*, *6*(2), pp. 226–249. https://doi.org/10.1558/jocr.v6i2.226

Elder-Vass D. (2010) *The Causal Power of Social Structures: Emergence, Structure and Agency*. Cambridge University Press.

Epstein B. (2015) *The ant trap: rebuilding the foundations of the social sciences*. Oxford University Press.

Epstein B. (2016) A Framework for Social Ontology. *Philosophy of the Social Sciences*, *46*(2), pp. 147–167. https://doi.org/10.1177/0048393115613494

Epstein B. (2018) Social Ontology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2018). Metaphysics Research Lab, Stanford University. https://plato.

stanford.edu/archives/sum2018/entries/social-ontology/

Fararo T. J. (1989) *The meaning of general theoretical sociology: tradition and formalization*. Cambridge University Press.

Gallotti M. (2012) A Naturalistic Argument for the Irreducibility of Collective Intentionality. *Philosophy of the Social Sciences*, *42*(1), pp. 3–30. https://doi.org/10.1177/0048393111426864

Gilbert M. (1992) *On Social Facts*. Princeton University Press.

Gintis H. (2007a) A framework for the unification of the behavioral sciences. *Behavioral and Brain Sciences*, *30*(1), pp. 1–16. https://doi.org/10.1017/S0140525X07000581

Gintis H. (2007b) The evolution of private property. *Journal of Economic Behavior & Organization*, *64*(1), pp. 1–16. https://doi.org/10.1016/j.jebo.2006.02.002

Gintis H. (2009) *The bounds of reason: game theory and the unification of the behavioral sciences*. Princeton University Press.

Guala F. (2007) The Philosophy of Social Science: Metaphysical and Empirical. *Philosophy Compass*, *2*(6), pp. 954–980. https://doi.org/10.1111/j.1747-9991.2007.00095.x

Guala F. (2016) Epstein on Anchors and Grounds. *Journal of Social Ontology*, *2*(1), pp. 135–147. https://doi.org/10.1515/jso-2016-0003

Guala F., & Hindriks F. (2015) A Unified Social Ontology. *The Philosophical Quarterly*, *65*(259), pp. 177–201. https://doi.org/10.1093/pq/pqu072

Guala F., & Mittone L. (2010) How history and convention create norms: An experimental study. *Journal of Economic Psychology*, *31*(4), pp. 749–756. https://doi.org/10.1016/j.joep.2010.05.009

Hacking I. (1999) *The Social Construction of What?* Harvard University Press.

Haslanger S. (2018) What is a social practice? *Royal Institute of Philosophy Supplements*, *82*, pp. 231–247.

Haslanger S. A. (2012) *Resisting reality: social construction and social critique*. Oxford University Press.

Hédoin C. (2016) Toward a Naturalistic Philosophy of Institutions. *Œconomia. History, Methodology, Philosophy*, (6–3), Article 6-3. https://journals.openedition.org/oeconomia/2400

Hédoin C. (2021) The Beliefs-Rules-Equilibrium Account of Institutions: A Contribution to a Naturalistic Social Ontology. *Journal of Social Ontology*, *7*(1), pp. 73–96. https://doi.org/10.1515/jso-2020-0001

Hempel C. G. (1965) Empiricist criteria of cognitive significance: Problems and changes. In *Can theories be refuted? Essays on the duhem-quine thesis* (pp. 65–85). Springer.

Herrmann D. A., & Skyrms B. (2021) Invention and Evolution of Correlated Conventions. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/717161

Hindriks F. (2019) Norms that Make a Difference: Social Practices and Institutions. *Analyse Und Kritik*, *41*(1), pp. 125–145. Scopus. https://doi.org/10.1515/auk-2019-410109

Hindriks F., & Guala F. (2015) Understanding institutions: Replies to Aoki, Binmore, Hodgson, Searle, Smith, and Sugden. *Journal of Institutional Economics*, *11*(3), pp. 515–

522. Scopus. https://doi.org/10.1017/S1744137415000120

Hindriks F. (2005) *Rules & institutions: essays on meaning, speech acts and social ontology: essays over betekenis, taalhandelingen en sociale ontologie = Regels & instituties*. Haveka BV.

Hindriks F. (2009) Constitutive rules, language, and ontology. *Erkenntnis. An International Journal of Scientific Philosophy*, *71*(2), pp. 253–275.

Hindriks F. (2022) Institutions and their strength. *Economics & Philosophy*, *38*(3), pp. 354–371. https://doi.org/10.1017/S0266267121000195

Hindriks F., & Guala F. (2021) The functions of institutions: etiology and teleology. *Synthese*, *198*(3), pp. 2027–2043. https://doi.org/10.1007/s11229-019-02188-8

Kincaid H. (2024) A naturalist approach to social ontology. *Synthese*, *203*(1), 27. https://doi.org/10.1007/s11229-023-04432-8

Kitcher P. (n.d.) *Explanatory Unification and the Causal Structure of the World*.

Kitcher P. (1981) Explanatory Unification. *Philosophy of Science*, *48*(4), pp. 507–531. https://doi.org/10.1086/289019

Lauer R. (2019) Is Social Ontology Prior to Social Scientific Methodology? *Philosophy of the Social Sciences*, *49*(3), pp. 171–189. https://doi.org/10.1177/0048393119840328

Lewis D. (2008) *Convention: A Philosophical Study*. John Wiley & Sons. (Original work published 1969)

Little D. (2020) Social Ontology De-Dramatized. *Philosophy of the Social Sciences*, *51*(1), pp. 13–23. https://doi.org/10.1177/0048393120916145

Machamer P., Darden L., & Craver C. F. (2000) Thinking about Mechanisms. *Philosophy of Science*, *67*(1), pp. 1–25. https://doi.org/10.1086/392759

Maynard Smith J. (1982) *Evolution and the theory of games*. Cambridge University Press.

Metzger L. P. (2018) Evolution and correlated equilibrium. *Journal of Evolutionary Economics*, *28*(2), pp. 333–346. https://doi.org/10.1007/s00191-017-0539-z

North D. (1990) *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press.

O'Connor C. (2019) *The Origins of Unfairness: Social Categories and Cultural Evolution* (First edition). Oxford University Press.

Parsons T. (1935) The Place of Ultimate Values in Sociological Theory. *The International Journal of Ethics*, 3(45). https://doi.org/10.1086/intejethi.45.3.2378271

Paternotte C. (2020) Social evolution and the individual-as-maximising-agent analogy. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *79*, 101225. https://doi.org/10.1016/j.shpsc.2019.101225

Rabinowicz W. (2018) Are Institutions Rules in Equilibrium? Comments on Guala's *Understanding Institutions*. *Philosophy of the Social Sciences*, *48*(6), pp. 569–584. https://doi.org/10.1177/0048393118798643

Rakoczy H., & Tomasello M. (2007) The Ontogeny of Social Ontology: Steps to Shared Intentionality and Status Functions. In S. L. Tsohatzidis (Ed.), *Intentional Acts and Institutional Facts: Essays on John Searle's Social Ontology* (pp. 113–137). Springer Netherlands. https://doi.org/10.1007/978-1-4020-6104-2_5

Ross D. (2012, August 6) *Coordination and the Foundations of Social Intelligence*. The Oxford Handbook of Philosophy of Social Science. https://doi.org/10.1093/oxfordhb/9780195392753.013.0020

Ross D. (2023) Scientific metaphysics and social science. *Synthese*, *202*. https://doi.org/10.1007/s11229-023-04358-1

Roversi C. (2021) In defence of constitutive rules. *Synthese. An International Journal for Epistemology, Methodology and Philosophy of Science*, *199*(5), pp. 14349–14370.

Ruben D.-H. (1989) Realism in the social sciences. In H. Lawson & L. Appignanesi (Eds), *Dismantling truth: Reality in the post-modern world* (pp. 58–75). Weidenfeld and Nicolson.

Sarkia M., & Kaidesoja T. (2023) Two approaches to naturalistic social ontology. *Synthese*, *201*(3), 104. https://doi.org/10.1007/s11229-023-04105-6

Saunders D. (2022) How to Put the Cart Behind the Horse in the Cultural Evolution of Gender. *Philosophy of the Social Sciences*, *52*(1–2), pp. 81–102. https://doi.org/10.1177/00483931211049770

Schelling T. C. (1980) *The Strategy of Conflict: With a New Preface by the Author*. Harvard University Press.

Schmid H. B. (2023) *We, Together: The Social Ontology of Us* (1st edn). Oxford University Press, New York. https://doi.org/10.1093/oso/9780197563724.001.0001

Searle J. (1995) *The Construction of Social Reality*. Simon and Schuster.

Searle J. (2010) *Making the Social World: The Structure of Human Civilization*. Oxford University Press.

Seitz F. (2020) Argumentation Evolved: But How? Coevolution of Coordinated Group Behavior and Reasoning. *Argumentation*, *34*(2), pp. 237–260. https://doi.org/10.1007/s10503-020-09510-6

Sellars W. (1962) Philosophy and the scientific image of man. *Frontiers of Science and Philosophy*, *1*, pp. 35–78.

Skyrms B. (2014) *Evolution of the social contract* (Second edition). Cambridge University Press.

Smith J. M., & Price G. R. (1973) The logic of animal conflict. *Nature*, *246*(5427), pp. 15–18.

Smith J. M., & Parker G. A. (1976) The logic of asymmetric contests. *Animal Behaviour*, *24*(1), pp. 159–175. https://doi.org/10.1016/S0003-3472(76)80110-8

Sterelny K. (2003) *Thought in a hostile world: the evolution of human cognition*. Blackwell.

Sterelny K. (2012a) From fitness to utility. In K. Binmore & S. Okasha (Eds), *Evolution and Rationality* (pp. 246–273). Cambridge University Press. https://doi.org/10.1017/CBO9780511792601.012

Sterelny K. (2012b) *The evolved apprentice: how evolution made humans unique*. The MIT Press.

Sterelny K. (2016) Cooperation, Culture, and Conflict. *The British Journal for the Philosophy of Science*, *67*(1), pp. 31–58. https://doi.org/10.1093/bjps/axu024

Sterelny K. (2021) *The Pleistocene Social Contract: Culture and Cooperation in Human Evolution* (1st edition). Oxford University Press.

Sugden R., & Zamarrón I. E. (2006) Finding the key: The riddle of focal points. *Journal of Economic Psychology, 27*(5), pp. 609–621. https://doi.org/10.1016/j.joep.2006.04.003

Thomasson A. L. (2003) Realism and human kinds. *Philosophy and Phenomenological Research, 67*(3), pp. 580–609. https://doi.org/10.1111/j.1933-1592.2003.tb00309.x

Tuomela R. (2002) *The Philosophy of Social Practices: A Collective Acceptance View* (1st edn). Cambridge University Press. https://doi.org/10.1017/CBO9780511487446

Tuomela R. (2013) *Social Ontology: Collective Intentionality and Group Agents*. Oxford University Press.

Ullmann-Margalit E. (1977) *The emergence of norms*. Clarendon Press.

Vanderschraaf P. (1998) Knowledge, Equilibrium and Convention. *Erkenntnis, 49*(3), pp. 337–369. https://doi.org/10.1023/A:1005461514200

Vlerick M. (2016) Explaining Universal Social Institutions: A Game-Theoretic Approach. *Topoi, 35*(1), pp. 291–300. https://doi.org/10.1007/s11245-014-9294-z

Vlerick M. (2019) The Evolution of Social Contracts. *Journal of Social Ontology, 5*(2), pp. 181–203. https://doi.org/10.1515/jso-2019-0041

Weber M. (1924) *Gesammelte Aufsätze zur Soziologie und Sozialpolitik*. Mohr.

Wright L. (1973) Functions. *The Philosophical Review, 82*(2), 139. https://doi.org/10.2307/2183766

Об авторе / About the author

Шевченко Валерий— аспирант Высшей Школы Экономики. Научные интересы: философия социальных наук, философия когнитивных наук, социальная онтология, теория игр.

https://orcid.org/0000-0001-5040-8897. E-mail: valerii.s.shevchenko@gmail.com

Valerii Shevchenko — PhD student at the Higher School of Economics. Research interests: philosophy of social science, philosophy of cognitive science, social ontology, game theory.

https://orcid.org/0000-0001-5040-8897. E-mail: valerii.s.shevchenko@gmail.com