Evolutionary stable correlation as a core problem of social ontology

Valerii Shevchenko

2023

Table of Contents

Evolutionary stable correlation as a core problem of social ontology

Introduction

In this paper, I argue that the emergence of evolutionary stable correlation is the core issue of naturalistic social ontology. According to rules-in-equilibria theory, social institutions are the central unit of social ontology (Guala, 2016), and coordination is its main mechanism rooted in evolution (Shevchenko, 2023). As institutions are normatively-driven self-sustaining behavioral regularities designed to solve coordination problems (Aoki, 2007; Lewis, 1969), they share many features with 'animal conventions' that help animals solve coordination problems and maintain stable relationships (Hindriks & Guala, 2015). Consequently, understanding the emergence of social institutions requires an examination of the evolutionary mechanisms that enable correlation of strategies with normative force as a key characteristic.

To expand, let us first look at Guala's (2016) argument that has the following logic:

1. social institutions are backed not by constitutive rules of the form "X counts as Y in (the context of) C", like in Searle (1995),[1] but

[1] Coordination equilibrium is a concept defined by philosopher David Lewis which

by regulative rules of the form "do X if Y"
2. from a game-theoretic point of view, regulative rules can be seen as agents' strategies that comprise a *correlated equilibrium*[2]
3. constitutive rules are linguistically transformed regulative rules with added theoretical term that represents a certain equilibrium
4. at the same time, many animal species including baboons, lions, swallowtails, and others exhibit behavioral patterns describable in the form similar to correlated equilibrium (Maynard Smith, 1982)
5. despite the similarity of mathematical representation, the cases of 'animal conventions' and human social institutions differ in scope of actionable signals. Building on Sterelny (2003), Guala puts for forward an idea that humans can invent and follow new rules, whereas animals are bound to genetically inherited sets of behavioral responses
6. the arbitrariness of rules that humans can invent and follow is grounded in and ontologically depends on shared representations of a given community
7. put differently, the difference in scope of actionable signals between animals and humans can be explained by humans having social epistemology that grounds social ontology.

Although sound, this argument has an Achilles heel: the evolutionary roots of correlation of strategies as the basis of any self-sustaining social coordination, human or not, are still obscure and underdeveloped.

Guala and Hindriks base their account on Maynard Smith's, who does not use the notion of correlated eqiulibrium explicitly and discusses what he calls a *bourgeois equilibrium* — a situation in animal territorial behavior, when the most optimal strategy for an animal is to fight for a territory it "owns" or not fight otherwise. This game is represented in the matrix below.

states that when two or more individuals are engaged in a coordination game, they will naturally gravitate towards the same outcome, as this is the most rational choice. The idea is that each individual will tend to choose the same outcome because they can both benefit from it. This is in contrast with Nash equilibrium, where each individual must make a choice that maximizes their own payoff without considering the other's payoff.

[2]Correlated equilibrium is a general solution concept introduced by Aumann (1974, 1987). As opposed to the classic Nash equilibrium, where players choose their strategies independently, here players choose strategies based on a public signal the value of which they assess privately, thus coordinating their actions according to a given correlation device.

	Hawk	**Dove**	**Bourgeois**
Hawk	$\frac{(V-C)}{2}$, $\frac{(V-C)}{2}$	V, 0	$\frac{(V-C)}{2}$, $V-C$
Dove	0, V	$\frac{V}{2}$, $\frac{V}{2}$	0, $\frac{V}{2}$
Bourgeois	$C-V$, $\frac{(C-V)}{2}$	$\frac{(C-V)}{2}$, $C-V$	$\frac{(C-V)}{2}$, $\frac{(C-V)}{2}$

Table 1: A game-theoretic matrix for a "hawk-dove-bourgeois" game from Maynard Smith's book "Evolution and theory of games". In this game, two players (represented by rows and columns) can choose to be either a hawk (fight for resources), dove (submit and share resources), or bourgeois (submit only when opponent is also bourgeois). The payoffs are determined by the value of the resource (V) and the cost of fighting (C). The table shows the payoff for each player given their own strategy and their opponent's strategy.

Guala and Hindriks interpret bourgeois equilibrium as a correlated one. However, there are at least two interpretations of it: *correlated equilibrium* and *evolutionary stable strategy* (ESS)[3] based on uncorrelated asymmetry. They are mathematically distinct, and we will look at both in detail later.

The presented ambiguity creates tension at the backbone of Guala's argument. It means that:

- either 'animal conventions' are mathematically different from human social institutions, for they represent ESS and not correlated equilibrium, and there comes the burden of showing how the former becomes the latter in the course of evolution;
- or that 'animal conventions' are themselves correlated, and there comes the burden of showing how humans acquired the capacity for social epistemology that ontologically grounds social ontology as rules-in-equilibria.

Taking into account the wealth of research on transition from ESS to correlation in game theory (Herrmann & Skyrms, 2021; Kim & Wong, 2017; Lee-Penagos, 2016; Metzger, 2018; Skyrms, 1994), the first option in resolving the tension in Guala's argument becomes insufficient. The transition from ESS to correlation does not intrinsically presuppose the emergence of intentional compliance to norms, as in social institutions,

[3]An ESS is a strategy which, if adopted by a population, is resilient to invasion by any alternative strategy. Mathematically, an ESS can be defined as a strategy profile $s = (s_1, s_2, ..., s_n)$ such that $\forall s^{'} \neq s$, we have $\pi(s, s) > \pi(s, s^{'})$, where π is the average payoff of the population playing the strategies s and $s^{'}$ (Maynard Smith, 1982).

which are normatively-driven and at the same time arbitrary, as will be covered later. Consequently, it will be needed to account for the second option, but to begin, we need to figure out whether social institutions indeed necessitate correlation of strategies. In this paper, I will address the source of the issue—Maynard Smith's notion of bourgeois equilibrium and its interpretations in regard to social coordination.

It is relevant, for if social institutions have emerged from 'animal conventions' with the aid of cognitive capacities like mindreading and/or mindshaping (Zawidzki, 2013), it constrains social ontology as the scope of possible objects of study to the logical derivatives of social institutions and social coordination in general as discussed in Shevchenko (2023).

This paper is structured as follows. The first section is devoted to description of the rules-in-equilibria (RiE) theory of social institutions. In the second section, the relationship between social institutions, conventions, and norms is discussed. The third section examines the notion of representation as used in RiE. In the fourth section, the role of Hawk-Dove-Bourgeois game as correlation and as asymmetry of strategies is studied. The final section explores the source of randomization in correlation as the problem in social institutions as evolved correlated equilibria.

Let us start with the notion of social institutions, destructure it into norms and conventions, study their relations and gradually arrive at the issue of coordination either by correlation or by asymmetry of strategies.

1. Social institutions as rules-in-equilibria

Hindriks & Guala (2015) present a unified theory of social institutions as rules in equilibria represented symbolically by theoretical terms like "money" or "marriage". It bridges accounts of regulative rules, equlibria of strategic games and 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.

According to Guala (2016), institutions as rules-in-equilibria are normatively-driven behavioral regularities comprising correlated equilibria. "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 and population thereof. Other scholars pinpoint normative and self-sustaining nature of institutions. They are "humanly devised constraints that shape human interactions" (North, 1990), "norm-governed social practices" (Tuomela, 2013) and "self-sustaining salient behavioral patterns" (Aoki, 2007). It can be seen that 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 equlibrium-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 (Searle, 1995).

According to the authors, the rule-based account is insufficient, for it cannot explain why some rules are followed and others not. To address this issue, an equilibrium account is needed to show the strategic character of rule-following.

Hindriks and Guala illustrate this point by comparing the two paradigmatic games from game theory, which are prisoner's dilemma and stag hunt. Although mutual defection in the prisoner's dilemma is a Nash equlibrium (NE),[4] it is not a social institution, however, 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, what explains why some rules are followed and other not.

However, the notion of players' correlated strategies as an *explanans* of the stability of institutions is insufficient, as the authors point out,

[4]Nash equilibrium is a solution concept describing a strategy profile consisting of each player's best response to the other player's strategies where no one gains bigger payoff by deviating unilaterally.

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

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

Table 2: Prisoner's dilemma (left) and Stag hunt (right)

for it is too permissive. The authors provide an example of non-human animals solving coordination problems but still not having institutions. For example, male baboons, lions, swallowtails and some other species exhibit a recurring behavioral pattern that can be described in terms of correlated equilibrium. 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" game to provide an example of a prototypical social institution:

	H	D	B
H	-1	2	0.5
D	0	1	0.5
B	-0.5	1.5	1.0

Table 3: small "Hawk-Dove-Bourgeois" game

Presented with a terrestrial resource, a "Hawk" player fights over it, a "Dove" retreats and "Bourgeois" uses a strategy "fight if own and retreat if do not own". In this game, Guala and Hindriks see the "bourgeois" strategy "fight of own" as a correlated one, meaning that players coordinate their actions by conditioning them on an external signal. As they say, its 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 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 most stable set of strategies is grazing if on the own side of the dry river. However, this is insufficient, for the payoff structure of the game is uniform for animal and human cases. Hence, we cannot discriminate between them solely on this basis.

	G	NG	GIS
G	-1	2	0.5
NG	0	1	0.5
GIN	-0.5	1.5	1.0

Table 4: Grazing game: the player strategies are Graze, Not Graze and Graze if North / Graze if South

What, then, distinguishes animal conventions from human social institutions? Guala and Hidriks argue that they differ in the scope of actionable signals. Building on the work of Sterelny (2003), they say that animals may only respond to a limited set of stimuli, but humans, with their ability to use representations and symbols to condition behavior, can decouple stimulus and response and invent new 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: establishing various correlations, devising new strategies, and expanding the number of equilibria.

There are two types of rules:

- agent-rules that agents formulate to represent and guide their behaviour
- observer-rules that observer formulates to represent and summarize others' behaviour.

Strategies can be described as rules, but a-rules influence behaviour 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

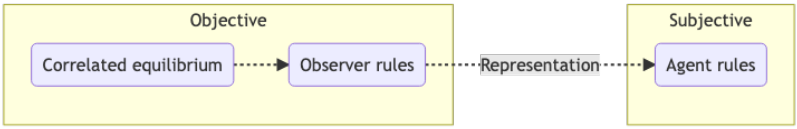
restrict certain behaviors and their own perpetuation. But how they are connected?

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

- coordination
- correlation
- representation.

Following the logic of the authors, institutions coordinate behavior by correlation of strategies, and humans are able to devise many strategies and equilibria given the same correlation device, or signal, due to an advanced cognitive capacity for conditioning behavior on representation of the environment. At the same time, as rules-in-equilibria theory has "rules" and "equilibria" parts, they must be somehow connected. For this reason, rules are symbolic representations of strategies in a game that comprise equilibria. 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".



However, it is not evident how exactly rules represent strategies. To clarify this issue, the authors, drawing on Hindriks (2005), propose to bridge their rules-in-equilbria account of institutions with the constitutive rules account. The latter presents institutions as systems of statuses and functions, paradigmatically proposed by Searle (1995) 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—for example, "Bills

issued by the Bureau of Engraving and Printing count as money in the United States" (Searle, 1995, p. 28). X 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).

$$\text{Regulative rules} + \text{Institutional terms} = \text{Constitutive rules}$$

For example, one can transform a regulative rule "if a bill is issued by the Bureau of Engraving and Printing, it can be used to pay for goods in the United States" into "Bills issued by the Bureau of Engraving and Printing count as money in the United States" by adding an institutional term "money". Now let us turn to the relationship between the concepts of rules, norms, conventions and institutions in rules-in-equilibria theory.

2. Rules, norms and conventions

What are the rules Guala and Hindriks are talking about? As they stipulate that institutions are norm-governed salient social practices, or behavioral regularities, rules are norms. It is the case for agent rules, though. So, what roles do norms play in institutions? Guala notes that social norms fulfill two functions highlighted by North (1990): they make behaviour more stable and more predictable. However, as noted by Searle, they introduce new behaviours, as well, and they do it by changing game payoffs. Norms help explain not only the persistence of institutions, but also its emergence. But if social norms are inherently important to institutions, what are they, and how do they differ from social institutions?

Hindriks (2019) elaborates on the definition of social institutions as norm-governed social practices and explicates how norms might govern

practices. His main idea is 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 the author, this 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 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.

Social norms can influence behavior due to sanctions imposed for violating them. Such sanctions modify people's preferences in cooperation games and motivate them to cooperate (Ullmann-Margalit, 1977). Apart from this, norms can be seen as legitimate, which leads people to conform even if it might not be in their best interest (Bicchieri, 2005). This is evidenced by the difficulty people experience when deciding to violate norms. In other words, decisions to conform are often more complex than a simple cost-benefit calculation.

Hindriks highlights coordination and cooperation types of social norms. Coordination norms such as first-come-first-serve as standing in line are contrasted by cooperation norms like "I-will-scratch-your-back-if-you-scratch-mine". Game-theoretically, this distinction is represented by either aligning or conflicting interests of agents in a game, respectively.

To this end, social institutions can be seen as solutions to coordination or cooperation games. Coordination games have at least two solutions that benefit all players. For example, two sides of the road to drive on. It does not matter on which particular side all the drivers drive, but the side being the same does matter, e.g, left in the UK or right in Europe. Cooperation games have one solution optimal for all players. For example, hunting a stag requires several participants but has a higher payoff for each, whereas hunting a rabbit can be done alone and has a lower payoff. It is more beneficial for everyone to cooperate and hunt a stag to get a higher payoff.

Hindriks studies the conditions of possibility for such behavioral regularities that successfully solve coordination and cooperation problems. He starts with the notion of convention, which is a population-wide beneficial regularity of behavior, deviating unilaterally from which is disadvantageous (Lewis, 1969). As there are two or more equally profitable solutions, or equilibria, in coordination problems, the mutual convergence of the agents on the same solution becomes important, oth-

erwise there will be miscoordination. Lewis, a pioneer of game-theoretic analysis of conventions, argues that given recurrent situations with coordination problems, people choose by precedent. They condition their behavior on what they expect others to do, enabling coordinated behavior among the population.

Lewis's account of conventions states that a behavioral regularity is a convention in a population P in a coordination game situation S if the following criteria are fulfilled:

- (1) Members of P conform to the regularity;
- (2) They expect others to conform;
- (3) They prefer to conform if others do so;
- (4) This is common knowledge of the form "everybody knows that everybody knows that P".

On this account, conventions are strict NE. At the same time, Lewis regards conventions as norms and does not make a sharp distinction between the two.

Vanderschraaf advances Lewis's notion of convention by formalizing the notion of salience central to Lewis's account. He shows that conventions are not NE, but correlated equilibria (Vanderschraaf, 1995, 2001).

On this game-theoretic account, social norms are modeled as sanctions with costs that can alter behavior by influencing agents' preferences, as agents face costs for not conforming to it. High costs and a greater likelihood of violation-detection increase the incentive to cooperate. Institutions, in their turn, are maintained partly because of these norm costs.

At the same time, introduction of sufficiently high delta parameter into cooperation problems transforms them into coordination ones. For instance, given a Prisoner's dilemma with a high delta parameter representing a cost for norm violation, the game becomes that of coordination with two equilibria — "Cooperate, cooperate" and "Defect, defect" (CC, DD) instead of only one — DD (Crawford & Ostrom, 1995). This shows that normative rules can in principle be coordination devices, or "choreographers", as Gintis puts it (Gintis, 2009).

Hindriks draws a distinction between social norms and conventions: norm-compliance is motivated, and conventions are self-reinforcing. He

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

	C	D
C	$-1,-1$	$-3,0-\delta$
D	$0-\delta,-3$	$-2,-2$

Table 5: Delta parameter transforming cooperation game into coordination game.

also calls them descriptive and normative conventions. Bicchieri (2005) has stated this in terms of the relationship between self-interest and common interest. They coincide in conventions and do not in norms. It means there can be conventions without norms. However, contra Lewis, Gilbert (1992) explicitly treats conventions as intrinsically normative and calls them quasi-agreements conceptually linked to joint intentions, which generate normative reasons for conformity. At the same time, Brennan et al. (2013) argue that conventions can become normative because they protect or promote some value. Guala & Mittone (2010) support this by empirical evidence.

Overall, conventions are self-reinforcing, so sanctions are not necessary for creating and maintaining a mutually beneficial behavioral regularity. However, both Lewis and Bicchieri acknowledge that exceptions may exist, making some conventions more fragile than others. Norms, in their turn, make confirming more attractive and thus help to stabilize conventions to ensure collective benefits and prevent malfunctions.

According to the rules-in-equilibria theory of institutions that Hindriks defend along with Guala, institutions are norm-governed social practices. And Hindriks defines a social practice as a regularity in behavior that involves norms. Practices arise in response to signaling devices, which are salient features of the environment that enable agents to align their behaviors in beneficial ways, creating new strategies and thus giving rise to conventions. Interdependent behavioral regularities in coordination games arise from signaling rules of the form “if D, do A”, whereby agents condition their behavior on a signal to coordinate mutually beneficial interactions and achieve collective benefits. In a case of a traffic light, the light itself serves a signaling device that helps to make traffic safer and more efficient by coordinating behaviors.

As normativity pervades social interaction, Hindriks distinguishes two types of normative standards: deontic ones, such as right/wrong, and evaluative ones, such as better/worse. Deontic standards signify oblig-

ations, while evaluative standards refer to the quality of performance in various activities, e.g., hosting guests. Social practices can feature either deontic and evaluative standards, evaluative standards only, or neither. As an example, a group of friends that loathes rules may also dislike evaluative standards. This suggests that conventions involve signaling rules, but do not necessarily include normative rules. Therefore, social practices can exist without being an institution.

Hindriks discusses several views of social norms. The 'normative-beliefs view' holds that when people encounter a coordination or cooperation game situation, they are expected to act in a certain way and this is generally known. Brennan et al. (2013) define social norms as normative principles or rules which are commonly accepted and known. Social norms are thus generally accepted and recognized normative beliefs.

However, the phenomenon of "pluralistic ignorance" counteracts the 'normative-beliefs view' by being too restrictive in requiring acceptance of the norm. Hindriks provides an example of college students believing that they are expected to drink heavily on weekends, while not really liking doing it. They do not believe that they ought to do so, but they acknowledge that others believe that college students generally do it. To reflect it, the secind, normative-expectations view, proposes that a social norm exists in a population if its rule is present in the normative expectations of its members. This differs from the normative-beliefs view in that it does not require acceptance of the norm, only acknowledgment of it, and that knowledge of others' attitudes is not necessary. It permits inclusion of the norm in first-order beliefs.

Bicchieri's theory (2005) is largely akin to the normative-expectations view, yet there are three key differences. Firstly, she limits the concept of a social norm to regulations that address cooperation problems, while Hindriks includes coordination ones, as well. Secondly, her conception of normative expectations does not make them normative, strictly speaking, for they are higher-order empirical expectations. Someone has a normative expectation if they expect others to adhere to a descriptive rule of 'Everybody does A'. According to Bicchieri, this involves obligations. However, as Hindriks stipulates, an expectation of behavior differs from the belief that a normative rule applies; the former being an expectation, the latter a belief about what should be done.

The third view of social norms is 'conditional-preferences view'. It holds

that a social norm exists when enough participants prefer to conform to it given empirical and normative expectations (Bicchieri, 2005). However, Southwood (2019) argues that people may secretly wish to break the norm if others do the same and expect each other to do so. According to the conditional-preferences view, perceiving a social norm as legitimate is when someone regards the relevant normative expectations as well-founded. This can motivate people to act accordingly (Bicchieri, 2005).

Overall, the normative-beliefs view holds that people need only possess normative beliefs featuring a rule for it to be perceived as legitimate; these beliefs are self-justifying. The conditional-preferences view, however, states that legitimacy is derived from justified normative expectations.

According to Hindriks, neither of these two views is adequate. The normative-beliefs view holds that social norms are self-justifying and tend to be regarded as legitimate. Yet, pluralistic ignorance shows that this is not always the case. The normative-expectations view suggests that perceived legitimacy is based on justified normative expectations, which lead to corresponding beliefs. This belief lies at the core of what it means for a social norm to be seen as legitimate, and can only be suitably justified with empirical and normative expectations. The conditional-preferences view fails to capture this complexity, while the normative-expectations view does so by explicating legitimacy in terms of an agent's normative beliefs. This motivates agents to conform and makes a difference to behavior within institutions.

Moreover, the normative-expectations view states that a social norm has authority if the normative beliefs people have are suitably justified. This is only true if the expectations are both justified and true, indicating that there is an applicable regularity and that others believe the norm applies.

According to Hindriks, for a social norm to govern a social practive, its participants must adhere to it. This will create an institution, which will be perceived as legitimate and have authority. However, norm-following alone is too demanding an explanation for institutions that are not seen as legitimate. Sanctions, which are important in formal and informal norms, demonstrate that norm-following does not always lead to conformity.

In sum, a norm governs a practice only if it motivates a substantial number of participants. It happens when it is deemed significant to conform to it. Norm-conformity is not enough for norm-governance, as demonstrated by the example of the convention to drive on the right-hand side of the road. This convention is self-reinforcing, but does not motivate anybody and does not constitute an institution. Thus, neither norm-following nor norm-conformity is necessary for norm-governance and norm-conformity is insufficient.

To this end, social institutions are norm-driven conventions, or social practices, that require cognitive capacities for recognition, complying to and changing of social norms.

3. The problem with "representation"

As might be seen from the exposition, the authors base their argument on the notion of insufficiency—of both rules and equilibria as distinct explanations of institutions. However, while justifying the insufficiency of equilibria with applicability of the concept of correlated equilibria to both humans and animals, the authors 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. It means that coordination and correlation are insufficient, and representation is needed.

However, the character of the term "representation" is ambiguous: a-rules "represent" game-theoretic strategies in a more philosophical sense and not in a cognitive one, while the authors mention terms like stimuli, behavior and representation, that clearly imply a narrower cognitive perspective. From a social-scientific point of view, representation as a relation makes sense, for it allows investigation of Y-terms, or institutional terms, used by agents by observing social practices, circumscribing social norms that govern them and then trying to figure out the respective strategies in equilibrium (Guala, 2016, ch. 14). However, representation as a cognitive capacity does not have any immediate practical application, especially in sociological data. Hence, there is need to discern two notions of representation in Guala's and Hindriks' argument:

- representation as relation
- representation as cognitive capacity.

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, it means that the representation as a relation between the rules and equilibrium might ontologically depend on representation as a cognitive capacity.

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 the 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, they mainly mean "representation-as-relation", as 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). According to the authors, 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 the 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).

Here is a problem with this argument. It presupposes that behavior conditioning, and hence strategy selection, occurs already being based on existing rules. To follow a rule, it should exist. 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 the existing strategies, and this involves a vicious circle: inventing new rules requires not only a capacity for stimulus-behavior decoupling, but existing equilbria, for here salient features of existing equilibria are used as coordination devices. In other words, the authors equate representation of salient features of the environment with representation of existing strategies, or behavioral responses, that

preexist 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. 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.

Would this work without representation as a cognitive capacity? 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 given the same coordination device, or signal. For example, if butterflies can coordinate only on the precedence of occupying the sun spot, for they use the temporal order of territory occupation as a coordination device, humans are not genetically hardwired for using one and only coordination device for a given situation. We can interpret the same coordination device differently. As a simple 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. And it crashes when a foreigner tries to understand a native. 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? No, as well, as it is the crux of the argument and of the unification done by rules-in-equilibria theory. Representation here logically connects rules and equilibria and helps to further connect it to constitutive rules theory by the notion of institutional, 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. 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. And if there is no decoupling, hence there is no "new rules

and strategies". 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" to introduce brand 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 whole argument about social institution as rules-in-equilibria to succeed, Guala and Hindriks should show two things:

- that correlated equilibrium is indeed supported both in human and animal conventions in the first place. It is for Maynard Smith (1982), from whom they take the notion of bourgeois equilibrium, uses ESS and not correlated equilibrium;
- that representation as a relation between rules and equilibrium is ontologically dependent on representation as a cognitive capacity.

For the theory to fully work, it is needed 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, if they do at all. However, this is out of scope of this paper. Now we take a step back and analyze the notion of a "Hawk-Dove-Bourgeois" game as introduced by Maynard Smith (1982) that plays a crucial role in the argument of Guala and Hindriks.

4. Correlation and asymmetry of strategies

Guala and Hindriks draw inspiration for their rules-in-equilibria theory of social institutions in Maynard Smith's concept of *"bourgeois equilibrium"* (Maynard Smith, 1982). They see the "Hawk-Dove-Bourgeois" (HDB) game of animal territorial ownership as representing a prototypical "animal convention". According to the authors, bourgeois equilibrium (BE) is essentially a correlated equilibrium (CE), however Maynard Smith uses bourgeois to define evolutionary stable strategies 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 it is unclear how they can be combined. Hence, the issue consists of clarifying the status of BE: whether its situation describes 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. Let us look at the Maynard Smith's notion of BE captured in the the HDB game.

4.1 Maynard Smith's "Hawk-Dove-Bourgeois" game

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; e.g., growth form, age at first reproduction, or relative number of offspring

He proposes 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	$\frac{1}{2}(V-C)$, $\frac{(V-C)}{2}$	V
Dove	0	$\frac{V}{2}$

Table 6: 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.

As this model is at the core of Guala's theory, its assumptions are important. This model assumes an infinite population with asexual reproduction and symmetric contests between two opponents. It also has a finite set of strategies.

Defining the stability criteria for the strategies, he proposes 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 $E(I,I) > E(J,I)$ or $E(I,I) = E(J,I)$
- 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

$$E(A,I) = E(B,I) = E(C,I) = ... = E(I,I)$$

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:

1. Contests are between two individuals of distinct roles (e.g., owner/intruder, larger/smaller, older/younger);
2. Both individuals know their role;
3. Both have the same strategies available (e.g., escalate, retaliate, display);

4. 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

Table 7: 'Hawk-Dove-Bourgeois' game

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.

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 $V < C$.[5]

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

[5]However, it is still not clear whether human players such as grazers have genuine fitness rather utility value function. As Sterelny (2012) 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.

not entail that she always plays a certain 'owner' strategy like 'Hawk' or 'Bourgeois'. However, according to Gintis (2007), 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.

Smith & Parker (1976) used the term 'uncorrelated asymmetry' to refer to contests in which 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. However, the act of punishing carries a cost, so in order to maintain stability, refusal to participate in enforcement must be considered a breach and punished as well. To ensure enforcement, a subgroup may be rewarded for carrying it out. ***This is essentially a 'Driving game' or any other Hi-Lo coordination game***

Overall, BE assumes that each player is trying to maximize their own self-interest, but no player is attempting to dominate or exploit the others. A BE is certainly a situation and not a solution concept. It occurs when the players have reached a strategy profile in which none of them can improve their payoff by changing only their own strategy, while also recognizing the other player strategies.

As there are two possible interpretation of BE—correlation of strategies and uncorrelated asymmetry, let us consider both.

4.2 Interpretation of HDB: correlation of strategies

Guala and Hindriks put forward that coordination in social institutions, and in 'Hawk-Dove-Bourgeois' 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 the concept of correlated equilibrium (CE) that goes beyond the Nash equilibrium and allows players to coordinate their strategies through the use of a common randomizer, such as a coin toss or a dice roll. This allows players to make decisions based on their beliefs about how the other players will act, which can increase the efficiency of their 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.

Formally, a correlated equilibrium is a probability distribution p over the set of action vectors S if the strategy vector τ^* is a Nash equilibrium of the game $\Gamma^*(p)$ (Zamir et al., 2013, p. 307). In other words, for every player $i \in N$:

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

The equation states that the optimal strategy for each player is dependent on both their own decisions and on those of other players, which reflects how CE allows players to take into account each other's behavior when making decisions.

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 $\Gamma^1 \times \cdots \times \Gamma^n$, 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.

Back to the HDB game, its important feature is that if $V < C$, the B strategy helps to settle contests *conventionally*. Maynard Smith does not emphasize the notion of convention, but it is key in Guala's discussion of social institutions as it describes mutually beneficial behavioral

regularities. What Maynard Smith means by conventional settlement is that there is shared 'understanding' of the situation between the players that helps to decide on the action. But what precisely does 'conventional settlement' mean regarding the B strategy in the HDB game? Let us start with convention as CE.

Vanderschraaf (1995) formalizes Lewis's notion of salience in coordination games and models conventions as correlated equilibria instead of Nash ones.

Lewis (1969), building on the ideas of Schelling (1980), proposed the notion of salience as an explanation of how a convention become established. A coordination equilibrium[6] of a game is salient if it is noticeable to all players, and they expect their opponents to choose the same equilibrium, resulting in them playing it. As Lewis suggests, salience can be determined by environmental factors.

Lewis considers a coordination equilibrium a convention if the players have common knowledge of a mutual expectations criterion (MEC). It means that each agent has a decisive reason to conform to her part of the convention, expecting the other agents to do likewise. He states that an equilibrium must be a coordination equilibrium to reflect the notion that a person conforming to a convention wants their intention to be seen as such. Vanderschraaf calls it the public intentions criterion (PIC). Furthermore, Lewis argues that common knowledge of the MEC is necessary for a convention. However, as Vanderschraaf notes, it is not sufficient, since common knowledge of the MEC can be satisfied at any strict Nash equilibrium.

Vanderschraaf defines a convention as a mapping of "states of the world" to strategy combinations of a noncooperative game (Vanderschraaf, 1995, p. 69):

DEFINITION 1. A *game* Γ is an ordered triple $(N, S, \mathbf{u})$ consisting of the following elements:

1. A finite set $N = \{1, 2, ..., n\}$, called the *set of players*;

[6]Coordination equilibrium is a concept defined by philosopher David Lewis which states that when two or more individuals are engaged in a coordination game, they will naturally gravitate towards the same outcome, as this is the most rational choice. The idea is that each individual will tend to choose the same outcome because they can both benefit from it. This is in contrast with Nash equilibrium, where each individual must make a choice that maximizes their own payoff without considering the other's payoff.

2. For each player $k \in N$, there is a finite set $S_k = \{A_{k_1}, A_{k_2}, \dots, A_{kn_k}\}$, called the *alternative pure strategies* for player k. The Cartesian product $S = S_1 \times \cdots \times S_n$ is called the *pure strategy set* for the game Γ;

3. A map $\mathbf{u} : S \to \mathbb{R}^n$, called the *payoff function* on the pure strategy set. At each strategy combination $\mathbf{A} = (A_{1j_1}, \dots, A_{nj_n}) \in S$, player k's payoff is given by the kth component of the value of $\mathbf{u}$, that is, player k's payoff u_k, at $\mathbf{A}$ is determined by

$$u_k(\mathbf{A}) = I_k \circ \mathbf{u}(A_{1j_1}, \dots, A_{nj_n}),$$

where $I_k(\mathbf{x})$ projects $\mathbf{x} \in \mathbb{R}^n$ onto its kth component.

As Vanderschraaf builds on Aumann's model (1987), each player has a personal *information partition* $\mathcal{H}_k$ of a probability space Ω. Elementary events on Ω are called *states of the world*. At each state ω, each player k knows which element $H_{kj} \in \mathcal{H}_k$ has occurred, but not which ω. H_kj represents k's private information about the states of the world. While k knows the opponent partitions, she does not know their content. A function $f : \Omega \to S$ defines a *exogenously correlated strategy n-tuple*, such that at each state of the world $\omega \in \Omega$, each player k selects a strategy combination $f(\omega) = (f_1(\omega), \dots, f_n(\omega)) \in S$ correlated with the state of the world ω. Thus, by playing $f_k(\omega)$, k follows *Bayesian rationality* and maximizes expected payoff given private information and expectations regarding opponents.

DEFINITION 2. Given $\Gamma = (N, S, \mathbf{u})$, Ω, and the information partitions $\mathcal{H}$ of Ω as defined above, $f : \Omega \to S$ is a *correlated equilibrium* if and only if, for each $k \in N$,

1. f_k is an $\mathcal{H}_k$-measurable function, that is, for each $H_{kj} \in \mathcal{H}_k$, $f_k(\omega)$ is constant for each $\omega' \in H_kj$, and

2. For each $\omega \in \Omega$,

$$E(u_k \circ f | \mathcal{H}_k)(\omega) \geq E(u_k \circ (f_{-k}, g_k) | \mathcal{H}_k)(\omega)$$

where E denotes expectation, '$-k$' refer to the result of excluding the kth component from an n-tuple. This holds for any $\mathcal{H}_k$-measurable function $g_k : \Omega \to S_k$. The correlated equilibrium f is *strict* if and only if the inequalities are all strict.

The measurability restriction on f_k means that k knows her strategy in each ω. This definition implies that players have common knowledge of the payoff structure, partitions of Ω, and $f : \Omega \to S$, which is needed to compute expected payoffs and reach correlated equilibrium. In addition, if the players possess common knowledge of Bayesian rationality, they will follow their ends of f, expecting others to do the same, since they jointly maximize expected utility in this way.

The agents refer to a common information partition of the states of the world. While each agent k has a private information partition $\mathcal{H}_k$ of Ω, there is a partition of Ω, namely the intersection $\mathcal{H} = \cap_{k \in N} \mathcal{H}k$, of the states of the world such that for each $\omega \in \Omega$, all the agents will know which cell $H(\omega) \in \mathcal{H}$ occurs. The agents' expected utilities in the following Definition 3 are conditional on their common partition $\mathcal{H}$, reflecting the intuition that conventions rely upon information that is public to all.

The agents' expected utilities are conditioned on their common information common partition $\mathcal{H}$ of the states of the world, which is the intersection of all their private partitions $\mathcal{H} = \cap_{k \in N} \mathcal{H}_k$. This reflects that conventions depend on information available to all agents.

DEFINITION 3. Given $\Gamma = (N, S, \mathbf{u}), \Omega$, and the partition $\mathcal{H}$ of Ω of events that are common knowledge among the players, a function $f : \Omega \to S$ is a convention if and only if for each $\omega \in \Omega$, and for each $k \in N$, f_k is $\mathcal{H}$-measurable and

$$E\left(u_k \circ f \mid \mathcal{H}\right)(\omega) > E\left(u_k \circ \left(f_{-j}, g_j\right) \mid \mathcal{H}\right)(\omega)$$

^6afd45

for each $j \in N$ and for any $\mathcal{H}$-measurable function $g_j : \Omega \to S_j$.

It means that if any player j deviates from a convention f, every player $k \in N$, including j, will be worse off. This definition of convention as a strict correlated equilibrium satisfies the PIC, as all agents are aware of the common partition and the strategies each player is expected to play. Thus, if any opponent mistakenly thinks that a player k will play a strategy $g_k(\omega) \neq f_k(\omega)$ other than the one prescribed by f, they may be tempted to deviate, resulting in a worse-off outcome for k. Conversely, if all opponents are aware that k will play her strategy $f_k(\omega)$ at each

state of the world $\omega \in \Omega$, then they have a strong incentive to conform with convention $f(\omega)$, which gives k an improved outcome.

Overall, Vanderschraaf's contribution is formalization of salience, hence he uses the *common* information partition $\mathcal{H}$ as a necessary restriction to make the definition of convention conform with Lewis' spirit. The other question is how salience itself emerges. Lewis suggests that pre-game communication, precedent, and environmental cues may lead agents to link their expectations and actions with various "states of the world", thus achieving correlated equilibrium. However, these sources of salience face the problem of infinite regress, for it is unclear how precedent or pre-game communication occurred in the first place without an established and shared conventional rules. Vanderschraaf, along with Skyrms (Vanderschraaf & Skyrms, 1993), proposes *inductive deliberation* as a mechanism by which salience is being established. It requires agents to be Bayesian rational and works by recursive belief modification. Players can reach a correlated equilibrium without communication by dynamically updating their beliefs using a common inductive rule, even if their beliefs don't initially allow for an equilibrium.

What is important in regard to the B strategy in the HDB game, Vanderchraaf notes that conventions as correlated equilibria allow for characterization of a wide range of equilibria. Given a game Γ with pure strategy coordination equilibria $\mathbf{A}_1, \ldots, \mathbf{A}_m, m \geqslant 2$, and a lottery Ω with mutually exclusive outcomes $H_1, \ldots, H_m$ such that $p_k\left(H_j = \lambda_j\right)$ for each player j. Then if the players condition on $\mathcal{H} = \{H_1, \ldots, H_m\}$, and $f : \Omega \to S$ is defined by $f(\omega) = \mathbf{A}_j$ if $\omega \in H_j$, then [[Convention is CE, as salience is an information partition#^6afd45|inequality]] is satisfied for all $\omega \in \Omega$, making f a convention. With infinitely many possible values for the λ_j's, any noncooperative game with two or more pure strategy coordination equilibria has infinitely many correlated equilibria corresponding to conventions.

Convention as correlated equilibrium allows for the "fair" coordination, even though no pure strategy equilibrium exists. Consider the "Battle of Sexes" game.

	$A1$	$A2$
$A1$	$10, 7$	$0, 0$
$A2$	$0, 0$	$7, 10$

Table 8: "Battle of sexes" game

Neither of the pure strategy Nash equilibria in this game is "fair", in the sense that the players receive the same payoff. This game has a mixed Nash equilibrium at which Player 1 plays $A1$ with probability $\frac{2}{3}$ and Player 2 plays $A2$ with probability $\frac{2}{3}$, and at this equilibrium each player's expected payoff is $\frac{2}{3}$, so this equilibrium is "fair". However, at the mixed Nash equilibrium, both players are indifferent to the strategies they play given what each player believes about her opponent, so this equilibrium fails the PIC and is consequently not a convention. Nevertheless, there is a correlated equilibrium fair to both players, and which each player will prefer over the pure strategy equilibrium that is unfair to her.

This game has a mixed Nash equilibrium at which both agents play their strategies with probability $\frac{2}{3}$, yielding an expected payoff of $\frac{2}{3}$ for each agent. However, this equilibrium does not satisfy the PIC and is thus not a convention. Nevertheless, there is a correlated equilibrium that is fair to both players and preferable to the pure strategy equilibrium.

With a toss of a fair coin, there is a probability space $\Omega = \{H, T\}$ with "heads" and "tails". The agents have a common information partition $\mathcal{H} = \{\{H\}, \{T\}\}$ and the correlated strategy combination is denoted as a function $f : \Omega \to \{A1, A2\} \times \{A1, A2\}$ with $f(H) = (A1, A1)$ and $f(T) = (A2, A2)$. Player 1 has a higher expected payoff with this combination than any of the other strategies, so she will not deviate from it. The expected payoff for Player 1 is 2 if the outcome is H, and 1 if it is T.

$$\begin{aligned} &E\left(u_1 \circ f \mid H\right) = 2 > 0 = E\left(u_1(A2, A1)\right) \mid H\right), \text{ and} \\ &E\left(u_1 \circ f \mid T\right) = 1 > 0 = E\left(u_1(A1, A2) \mid T\right) \end{aligned}$$

The same holds for the second player. To this end, neither player would want to deviate, since the overall expected payoff at this equilibrium for each player is

$$E\left(u_k \circ f\right) = \frac{1}{2} \cdot E\left(u_k \circ f \mid H\right) + \frac{1}{2} \cdot E\left(u_k \circ f \mid T\right) = \frac{3}{2}$$

It means that each player prefers the expected payoff from f to that of the mixed equilibrium.

One intrinsic problem with B as CE, however, is the source of randomization. Some scholars appeal to Nature as to a such source, calling it a *correlation device*, 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 who observes a random variable γ on a probability space (Γ, p) and issues directives $f_i(\gamma) \in S$ to each player i, such that choosing the directive is a best response given agents having a common prior p and assuming other players are also following Nature's directives (Gintis, 2009, pp. 135–136).

In their theory, Guala and Hindriks appeal to Skyrms's interpretation of the "Hawk-Dove" that is with correlation. According to Skyrms, the implicative nature of the B strategy is genuinely correlative. According to Skyrms, the (B, B) strategy profile is CE spontaneously arising from symmetry-breaking that happens when individuals randomize the choice of their strategies and do not know whether they are "Hawkes" or "Doves" (2014, p. 78).

However, ...

4.3 Interpretation of HDB: uncorrelated asymmetry

Another interpretation of HDB involves uncorrelated asymmetry instead of correlation. O'Connor (2019) employs this interpretation in her treatment of emergence of unfairness due to social categories as solutions to inherently institutional coordination problems. On this account, HDB strategy profiles are based not on correlation, but on uncorrelated asymmetry. It is a feature of games where players extract additional information from environment not included in the structure of a game. For example, they know that they are "Hawks" or "Doves" rather than their strategies are randomized. This underlies an important methodological distinction between correlated equilibrium and evolutionary stability.

5. Evolution, Bayesian updating and correlation

Conclusion

References

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

Aumann, R. J. (1974). Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, *1*(1), 67–96. https://doi.org/10.1016/0304-4068(74)90037-8

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

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

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

Brennan, G., Eriksson, L., Goodin, R. E., & Southwood, N. (Eds.). (2013). *Explaining norms* (1st ed). Oxford University Press.

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

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

Gilbert, M. (1992). *On Social Facts*. Princeton University Press. https://books.google.com?id=yYvcDwAAQBAJ

Gintis, H. (2007). The evolution of private property. *Journal of Economic Behavior & Organization*, *64*(1), 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. (2016). *Understanding institutions: The science and philosophy of living together*. Princeton University Press.

Guala, F., & Hindriks, F. (2015). A UNIFIED SOCIAL ONTOLOGY. *The Philosophical Quarterly*, *65*(259), 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), 749–756. https://doi.org/10.1016/j.joep.2010.05.009

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

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

Hindriks, F., & Guala, F. (2015). Institutions, rules, and equilibria: A unified theory*. *Journal of Institutional Economics*, *11*(3), 459–480. https://doi.org/10.1017/S1744137414000496

Kim, C., & Wong, K.-C. (2017). *Evolutionarily Stable Correlation*. *33*(1), 40.

Lee-Penagos, A. (2016). *Learning to coordinate: Co-evolution and correlated equilibrium* (Working Paper No. 2016-11). CeDEx Discussion Paper Series. https://www.econstor.eu/handle/10419/163012

Lewis, D. (1969). *Convention: A Philosophical Study*. John Wiley & Sons. https://books.google.com?id=GgCkLtTqBsMC

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), 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. (2015). The Place of Ultimate Values in Sociological Theory. *The International Journal of Ethics*. https://doi.org/10.1086/intejethi.45.3.2378271

Schelling, T. C. (1980). *The Strategy of Conflict: With a New Preface by the Author*. Harvard University Press. https://books.google.com?id=7RkL4Z8Yg5AC

Searle, J. (1995). *The Construction of Social Reality*. Simon and Schuster. https://books.google.com?id=zrLQwJCcoOsC

Shevchenko, V. (2023). Coordination as Naturalistic Social Ontology: Constraints and Explanation. *Philosophy of the Social Sciences*, 004839312211504. https://doi.org/10.1177/00483931221150486

Skyrms, B. (1994). Darwin Meets the Logic of Decision: Correlation in Evolutionary Game Theory. *Philosophy of Science*, *61*(4), 503–528. https://doi.org/10.1086/289819

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

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

Smith, J. M., & Price, G. R. (1973). The Logic of Animal Conflict. *Nature*, *246*(5427), 15–18. https://doi.org/10.1038/246015a0

Southwood, N. (2019). Laws as Conventional Norms. In D. Plunkett, S. Shapiro, & K. Toh (Eds.), *Legal Norms, Ethical Norms: New Essays on Meta-Ethics and Jurisprudence*. Oxford University Press.

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

Sterelny, K. (2012). 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

Tuomela, R. (2013). *Social Ontology: Collective Intentionality and Group Agents*. Oxford University Press. https://books.google.com?id=6ltpAgAAQBAJ

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

Vanderschraaf, P. (1995). Convention as correlated equilibrium. *Erkenntnis*, *42*(1), 65–87. https://doi.org/10.1007/BF01666812

Vanderschraaf, P. (2001). *Learning and coordination: Inductive deliberation, equilibrium, and convention*. Routledge.

Vanderschraaf, P., & Skyrms, B. (1993). Deliberational Correlated Equilibria. *Philosophical Topics*, *21*(1), 191–227. https://www.jstor.org/stable/43154147

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

Zamir, S., Maschler, M., & Solan, E. (2013). *Game theory*. Cambridge University Press.

Zawidzki, T. W. (2013). *Mindshaping: A New Framework for Un-*

derstanding Human Social Cognition. The MIT Press. https://doi.org/10.7551/mitpress/8441.001.0001