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), 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]For example, "bills issued by the Bureau of Engraving and Printing (X) count

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 of correlated equilibrium, as well (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" it or not fight otherwise. This game is represented in the matrix below.

Guala and Hindriks interpret bourgeois equilibrium as a correlated one. However, there are at least two interpretations of it: *correlated equi-*

as dollars (Y) in the United States (C)" (Searle, 1995, p. 28).

[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.

librium 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, which are normatively-driven and at the same time arbitrary, as will be covered later. Consequently, it will be needed to account for the

[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).

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. First, it discusses the relationship between social institutions, conventions, and norms, and how conventions emerge in Skyrms's dynamics and Harms's evolutionary functionalism. Second, it examines the correlation and asymmetry of strategies in the emergence of social institutions. Third, the paper explores the source of randomization in correlation as the problem in social institutions as evolved correlated equilibria. We will analyze Guala's argument about the difference in scope of actionable signals in animals versus humans and Skyrms's interpretation of Maynard Smith's "bourgeois" concept. Fourth, it delves into the tension between bourgeois and correlated equilibria with a formal distinction between mixed-strategy and 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.

Institutions vs. norms vs. conventions

According to Guala (2016), institutions are rules-in-equilibria, normatively-driven behavioral regularities represented as 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.

If social norms are inherently important to institutions, what are they, and how do they differ from social institutions? According to Bicchieri (2005), social norms are shared expectations, or "rules", about how people should behave in a given context. These expectations can be either prescriptive, telling individuals what they ought to do, or descriptive, reflecting what most people actually do. Social norms can be modeled as a set of rules or constraints that guide individual behavior. For example, let X be the set of all possible behaviors that an individual can choose from in a given situation. A social norm N can then be represented as a subset of X that specifies which behaviors are considered acceptable or desirable by the group: $N \subseteq X$. The power of social norms lies in their ability to shape behavior without the need for formal enforcement mechanisms like laws or explicit regulations. Individuals often conform to social norms because they want to fit in and be accepted by their peers, or because they believe that following the norm is the right thing to do. Thus, norms are shared expectations about behavior in certain situations and institutions are behavioral patterns that are governed by such shared expectations.

The further required distinction to be made is that of institutions, norms, and conventions. But what are conventions in the first place? Lewis (2008) defines conventions as regularities in behavior that are mutually expected and mutually beneficial for the agents involved. In other words, conventions are shared expectations about behavior that result in cooperative outcomes. To illustrate this concept, Lewis uses the example of driving on the right or left side of the road. This convention is mutually expected because everyone understands that it is necessary for traffic to flow smoothly and avoid accidents. It is also mutually beneficial because if everyone follows the convention, then there is a reduced risk of accidents and delays. Lewis also distinguishes between two types of conventions: coordination conventions and strategic conventions. Coordination conventions are those where agents need to coordinate their actions to achieve a common goal, such as deciding which side of the road to drive on. Strategic conventions are those where agents need to make strategic choices based on what they expect others to do, such as deciding whether to use a turn signal while

driving.

For example, consider the following coordination game:

	Drive on left	Drive on right
Drive on left	(1,1)	(-1,-1)
Drive on right	(-1,-1)	(1,1)

In this game, two drivers must choose whether to drive on the left or right side of the road. The payoffs indicate how well each driver does depending on their choice and their partner's choice. If both drivers choose the same side (either both drive on the left or both drive on the right), they each receive a payoff of 1. If they choose different sides (one drives on the left while the other drives on the right), they each receive a payoff of -1. This game has two pure strategy Nash equilibria:[4] both drivers driving on the left or both driving on the right. In other words, if both drivers follow these conventions, they will achieve a mutually beneficial outcome. Lewis argues that conventions can emerge in situations like this through repeated interactions between agents who learn to coordinate their behavior over time. As more people adopt a particular convention, it becomes more costly for others to deviate from it because they risk being penalized by their partners.

If conventions are mutually expected and mutually beneficial behavioral regularities, how are they different from both social norms and social institutions? O'Connor (2019) draws two crucial distinctions, namely between conventions and social norms, and between functional and arbitrary conventions. The former distinction implies that not all behavioral regularities possess normative force, meaning that conventions and norms are not the same. For instance, friends may have a convention of meeting every Friday evening at a bar, and failing to show up does not necessarily imply a violation of a norm. However, when two cars are driving in the same direction towards each other on the same side of the road, the drivers are compelled to swerve to avoid collision. Failing to do so may result in fines or even accidents; hence, swerving becomes an obligatory normative action.

Furthermore, as Bicchieri (2005) asserts, conventions differ from social norms in their association with self-interest and common interest.

[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.

While they converge with self-interest, they do not necessarily coincide with common interest. In the case of friends gathering at a bar, there is minimal or no tension between self-interest and common interest; however, when driving cars on the road, there is an inherent tension between these interests. O'Connor notes that conventions and norms exist along a continuum, where conventions can acquire normative force based on their position on this spectrum.

The second distinction pertains to the arbitrary and historically contingent nature of conventions, with the recognition that they are subject to variation and could have been otherwise. This arbitrariness is a fundamental characteristic of conventions, as posited by Lewis. However, Gilbert (1992) has critiqued Lewis's work, noting that not all potential resolutions to a coordination problem offer equal benefits for participants. Hence, where one mode of coordination is more desirable than another, conventionality is not entirely arbitrary. To put it differently, arbitrariness in the context of conventions illustrates a continuum ranging from necessity to contingency. For example, signaling among vervet monkeys may be construed as a convention in the Lewisian sense of recurrent behavioral patterns resolving coordination problems (cf. Harms, 2004; Skyrms, 2010). Nevertheless, this conventionality is not historically contingent insofar as multiple solutions are equally remunerative since adaptive dynamics breaks the symmetry between equilibria. Agents may be genetically predisposed towards certain strategies. Some conventions are more functional and others are more arbitrary.

Putting this into a perspective:

- 'animal conventions' are more functional conventions where "normativity", if exists, is grounded in genetically inherited behavioral predispositions;
- social institutions are more arbitrary conventions where normativity is grounded in advanced cognitive capacities like mindreading.

Essentially, social institutions are norm-driven conventions that require cognitive capacities which make recognition, complying to and changing of social norms possible. Two questions arise:

- if institutions are evolved 'animal conventions', how do the latter evolve themselves?
- do simple 'animal conventions' and social institutions evolve by

the same evolutionary mechanism?

Emergence of functional and arbitrary conventions

"Animal conventions" are behavioral regularities, where animals "know" how to behave. But how do they "know" that and how these regularities were established in the first place? Baraghith (2019) compares game-theoretic and teleosemantic views on emergence of conventions as public meaning. His main claim is that theories of Skyrms (2010) and Millikan (1987) share many aspects and can be synthesized to yield empirically testable and philosophically elaborated approach.

The author observes that signals, or public representations, become conventional by stabilization of a strategy profile in a Lewis signaling game, resulting in the emergence of behavioral regularities among involved agents. In other words, convention is generated by stabilization of a signaling system. According to Lewis, a signaling system is a *strict Nash equilibrium*[5] of a signaling game.

One of the similarities in teleosemantic and signaling approaches is that evolution drives the emergence of successful coordination between agents, be it parts of an organism or different organisms. However, teleosemantic approach operates with the notion of *function* (Millikan, 1987), whereas sender-receiver approach emphasizes *adaptive dynamics* by reinforcement learning (Skyrms, 2010).

In both approaches, conventions depend on their history and involve contingency. As Millikan (2005, p. 29) puts it:

> "A convention is merely a pattern of behavior that is (1) handed down from one person, pair, or group of persons to others – the pattern is reproduced – and (2) is such that, if *the pattern has a function*, then it is not the only pattern that might have served that function about as well. Thus, if a different precedent had been set instead, a different pattern of behavior would probably have been handed down instead."

[5]A strict Nash equilibrium is a Nash equilibrium where the player would even do worse by deviating unilaterally.

As Baraghith notes, most criticism of teleosemantic view of the emergence of conventions has been that content—or representation of a world state by a sender—lacks explanation solely by its adaptive function or history. However, as Neander and Shea show, teleosemantics might solve the problem of mental content, intentionality, and thus, representation (Neander, 2008; Shea, 2018). In its turn, sender-receiver approach has received criticism for being atomistic and not able to accommodate "mental life"—cases with agents having advanced cognitive capacities like midredaing (Sterelny, 2017).

Baraghith stresses the crucial difference between speaker meaning and public meaning. A convention as a signaling system involves two kinds of information: a representation of an observed world state by a sender, and a signal sent from sender to receiver. The former is internal, and the latter is external. Representation and signal are mental and behavioral parts of a representational system, respectively.

If a signaling system is a strategy profile of a signaling game, what is the latter? A signaling game represents a coordination problem between world states, signals and acts, which are associated probabilistically. The most simple case has two states $W = \{\sigma_1, \sigma_2\}$, two messages $M = \{m_1, m_2\}$ that a sender S can transfer to a receiver R, and two acts $A = \{\alpha_1, \alpha_2\}$, by which R can respond to a received signal. There are pure sender and receiver strategies. The former is a function $s : W \mapsto M$ from world states to signals, and the latter is a function $r : M \mapsto A$ from signals to acts. With two signals and two acts, both sender and receiver have 4 strategies each. Assuming that all strategies are equiprobable, 16 strategies are possible, from which only two are beneficial for both agents and constitute a strict NE.

In an evolutionary perspective of Skyrms (2010), signaling systems are not strict Nash equilibria, but ESS. On this account, given an adaptive process that guides the behavior of agents, any signaling game iterated over time results in an ESS. Depending on initial conditions, population converges on one of the two signaling systems, what is often modeled with replicator dynamics.[6]

[6]Replicator dynamics is a mathematical model used to describe the evolution of biological populations. It is based on the idea that individuals in a population can replicate themselves over multiple generations, and that their success or failure depends on their behavior relative to other members of the population. Mathematically, it is given by $\dot{x_i} = x_i(f_i(x) - \bar{f}(x))$, where x_i is the proportion of individuals in the population exhibiting a particular behavior, $f_i(x)$ is the fitness associated

Another detail of this approach is its connection to information theory. A signal m_1 carries information if it changes probabilities of a world state. The information quantity is measured by how far the probability is moved, and information content—by direction of probability: increasing or decreasing. Franke and Wagner (2014) show the Bayesian likelihood of a world state σ_i given a signal m_j:

$$P(\sigma_i \mid m_j) = \frac{P(m_j \mid \sigma_i) \times P(\sigma_i)}{\sum_t P(m_j \mid \sigma_i) \times P(\sigma_i)}$$

It means that if state σ_i occurs with prior probabilities $P(\sigma_i)$ and $P(m_j \mid \sigma_i) > 0$, signal m_j is sent. Signals may initially contain no intrinsic meaning, and the dynamics does not require any sophisticated cognitive capacities of the agents. They do not need to have pre-existing mental language for a signaling system to be established (Skyrms, 2010, p. 7). This makes sense in "animal conventions", but not easily so in human ones. As Huttegger puts it in regard to human language (2007, p. 413):

> "There is at least one functional aspect of human languages that can fundamentally be expressed in terms of signaling systems: communication facilitates social coordination".

However, it is not sufficient for evolutionary account of human social coordination resulting in social institutions.

Another important formal approach to the emergence of conventions is due to Harms (2004). He synthesizes sender-receiver framework and Millikan's teleosemantics. According to this approach, any semantic convention, or "rule", is a "function-stabilizing mechanism". It helps to coordinate the behavior of different organisms or different parts of an organism to perform an evolutionary adapted biological function. Rules are sets of maps from conditions to processes one by one. They say what to happen next given a state of the world. Rules for evolutionary adapted traits (AT) might be expressed as

$$R_{AT} = \{\langle c_{i,} p_i \rangle \mid AT\text{sel}p_i\text{in}c_i\}$$

A rule for an adaptive trait is a set of all ordered pairs of a condition and a process such that the trait was selected for performing the process p_i in the conditions c_i. (Harms, 2004, p. 203). Since any signaling system

with that behavior, and $\bar{f}(x)$ is the average fitness in the population.

consists of states, signals and acts, both signals and acts are adaptive. It means that a convention contains at least two rules:

- a rule of extension — it relates world states and signals by correspondence, has truth-value and is governed by a production mechanism P in a signal sender;

$$R(\text{extension})_P = \{\langle \sigma_i, m_j \rangle \mid P\text{sel}m_j\text{in}\sigma_i\}$$

- a rule of intension — it relates signals and acts by causal processes of interpretation and is governed by a response mechanism in a signal receiver;

$$R(\text{intension})_C = \{\langle m_i, \alpha_j \rangle \mid C\text{sel}\alpha_j\text{when}m_i\}$$

  Harms mentions a third rule, that of a signal production, but it is not relevant for us now:

$$R(\text{production})_P = \{\langle stimulus_i, m_j \rangle \mid P\text{sel}m_j\text{to}stimulus_i\}$$

  Rules $R(\text{extension}_P)$ and $R(\text{intension}_C)$ can be said to comprise a convention as a functional behavioral regularity, for both messages and acts are adapted — the former to world states and the latter to messages. But is it sufficient to define an animal convention?

It has been observed that animal signals not only inform about the world states, but also direct the behavior of others. For example, alarm calls of vervet monkeys both convey "Look, there is a leopard!" and "Run up the nearest tree" (Baraghith, 2019; Seyfarth & Cheney, 1990). Harms calls this "primitive content" that has both indicative and imperative functions (Harms, 2004, p. 189). Millikan calls it "pushmi-pullyu" representation and notes that purely descriptive and directive representations require a more advanced cognitive process than primitives (Millikan, 2005, p. 166).

Evolutionary development of primitive content leads to the divergence of its descriptive and directive functions due to advanced cognitive capacities. As Harms suggest, it introduces a stabilizing, or regulatory mechanism SM that works "atop" of conventions as rules for adaptive traits and guides behavior in case of failure of R_{AT}. It employs a corrective signal $CS = \{cs_1, ..., cs_n\}$ to "enforce" the initial convention

when a signal is not sent in the presence of a world state it was selected for:

$$R_{SM} = \{\langle \sigma_i \wedge \neg m_j \text{where} \langle \sigma_{i,} m_j \rangle \in R_{AT} \rangle \mid SM\text{sel}cs\text{when}(\sigma_i \wedge \neg m_j)\}$$

The rule for a stabilizing mechanism is a set of ordered pairs consisting of the failure of an adaptive trait and a corresponding corrective signal. If the adaptive trait fails, the stabilizing mechanism will detect this failure and send a corrective signal/action to restore it[7]. This division is echoed in Millikan's work as first-order and higher-order reproductive families (Millikan, 1987, p. 23). According to it, conventions R_{AT} are first-order and stabilizing mechanisms R_{SM} are second-order reproductive families that serve the same goal of restoring a first-order proper function.

It is tempting to say that on an O'Connor's "convention—social norm" continuum, R_{AT} is closer to conventions and R_{SM} is to norms, but in Harms conventions *are* function-stabilizing mechanisms containing normative component by definition. Hence, it means that convention contains both adaptive rules—for extension and intension—and their stabilizing mechanisms:

$$\text{convention} = \{R(\text{ex})_P, \quad R(\text{in})_C, \quad R_{SM}\}$$

$$\text{s.t.}.(Pselm_j \text{in} \sigma_i) \wedge (Csel\alpha_j \text{when} m_i) \wedge (SMselcs\text{when}(\sigma_i \wedge \neg m_j)\}$$

Thus, if a functional convention has normativity by default, and if institutions are norm-driven behavioral regularities, how do they differ from "animal conventions"? According to Guala and Hindriks, the difference is in scope of actionable signals. Animals have a more limited set of actionable signals than humans, as their behavior is tightly coupled to the stimuli by genetic wiring. In game-theoretic terms, for a signaling system is an ESS, it is disadvantageous for agents to deviate from the signal-act coupling. However, as Sterelny (2003) suggests, humans and other complex creatures have the ability to decouple stimulus and behavioral response with representations of environment. This, in theory, allows for invention and following different rules for the same signal. In formal terms it allows for several interpretations:

[7]There is an interesting similarity between a semantic regulatory mechanism like Harms' and regulatory networks in biology, that govern the dynamical repertoire of a given system like structural and regulatory genes [(Albert & Thakar, 2014)].

1. the cardinality of all the sets $|R(\text{ex})_P|$, $|R(\text{in})_C|$ and $|R_{SM}|$ is increased, which means that the capacity for richer behavioral response is dependent on the capacity to represent more world states, as Godfrey-Smith and Planer suggest (2014; 2021). It would imply that a convention is rich and complex and has many state-signal-act combinations as well as many stabilizing mechanisms restoring the function of these compinations;

```
graph LR;
w1(State 1)-->s1(Signal 1)-->a1(Act1)
w2(State 2)-->s2(Signal 2)-->a2(Act2)
```

2. the cardinality of only the intension rule $|R(\text{in})_C|$ is increased, meaning that the same set of world states can cause to produce different signals and acts. However, it is not clear what ensures the discrimination of states and successful production of different signals. Moreover, this configuration destroys the ordered pair structure of convention rendering it not functional anymore in the sense established earlier:

```
graph LR;
w1(State 1)--?-->s1(Signal 1)-->a1(Act1)
w1(State 1)--?-->s2(Signal 2)-->a3(Act2)
w1(State 1)--?-->s3(Signal 3)-->a4(Act3)
```

3. only the number of possible acts increases, hence the ordered pair structure of convention is displaced. This configuration is not functional, as well:

```
graph LR;

    w1(State 1)-->s1(Signal 1)
    s1--?-->a1(Act1)
    s1--?-->a2(Act 2)
    s1--?-->a3(Act 3)
    %%r1<-.Representation.->w1%%
    %%r1-->a2(Act2)%%
```

How to decouple stimuli as signals from behavioral responses as acts and preserve the structure of convention as a set of functional rules for extension, intension and stabilization? It leaves us with a tradeoff:

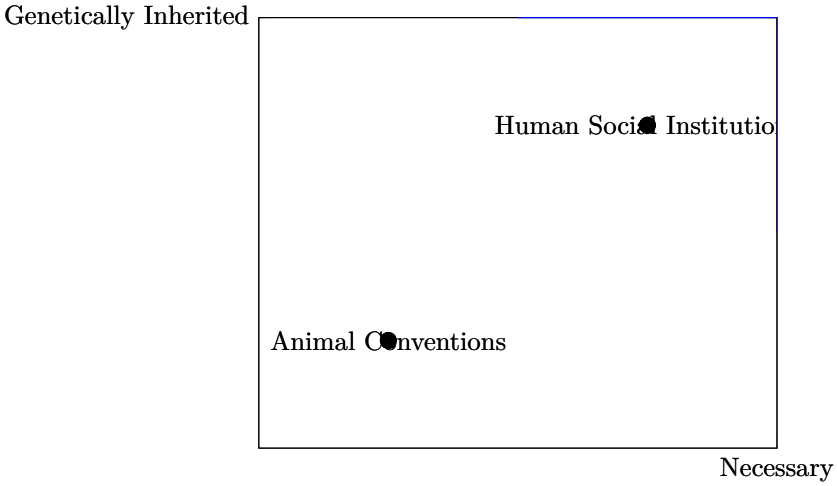
- either convention is indeed functional, but there is no "proper

decoupling", for there are just more ordered pairs of state-signal and signal-act;
- or human convention is not functional, and "proper decoupling" occurs by breaking the one-to-one mapping from signals to acts.

In O'Connor's terms, the transition from narrow set of actionable signals to the wider one is that from functional to arbitrary conventions. It means that the behavioral regularities "might have been otherwise", just like Guala and Hindriks propose. However, the game-theoretic implications of this are unclear, and several questions arise:

- do social norms as expectations evolve due to an increased degree of arbitrariness in conventions?
- what introduces arbitrariness into functional conventions? If, according to Guala and Hindriks, representations of environment are key differentiator for a wider set of actionable signals, are they what introduces arbitrariness?
- although not a game-theoretic question, but how does representation of environment itself evolve?

In other words, we need to study the relationship between the two axes of "convention space" built upon O'Connor's two distinctions.



O'Connor herself proposes a way to measure the degree of arbitrariness in conventions (2021).

Correlation and asymmetry of strategies

So far, we have established that conventions as "function-stabilizing mechanisms" might evolve from repeated signaling games and that it is possible to measure their arbitrariness. The next important question is what kind of equilibrium a convention is if it of evolutionary origin?

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" game of animal territorial ownership as a prototypical "animal convention". According to the authors, bourgeois equilibrium is essentially a correlated equilibrium, however Maynard Smith uses bourgeois to define ESS. It creates tension, for correlated equilibrium and ESS are mathematically distinct: the former is "too loose" and the latter is "too strict", and it is unclear how they can be combined. Hence, the issue consists of clarifying the status of bourgeois equilibrium in comparison to correlated equilibrium, ESS and mixed-strategy equilibrium, as well. It is due to being at the core of Guala's argument for institutions as correlation rooted in evolution. Let us look at the Maynard Smith's notion of bourgeois equilibrium represented by the "Hawk-dove-bourgeois" (HDB) game.

Maynard Smith's "Hawk-dove-bourgeois" game

Maynard Smith introduces the notion of "bourgeois equilibrium" (BE) in the context of animal behavior in evolutionary perspective (1982).

It is a game-theoretic solution concept that takes into account that players may not always be able to perfectly predict each other's moves and reach an ideal Nash equilibrium. Instead, they settle for a BE which is an acceptable compromise between their own and their opponents' goals. It assumes that each player is trying to maximize their own self-interest, but no player is attempting to dominate or exploit the others. A "bourgeois equilibrium" occurs when all players have reached a strategy profile (a combination of strategies for all players) such that none of them can improve their payoff by changing only their own strategy, while also recognizing the strategies of the other players. In BE, each player chooses a strategy independently. This is distinct from mixed-strategy equilibrium, correlated equilibrium and evolutionary stable correlation, which involve coordination or communication

among players.

More precisely, BE is a type of ESS where individuals cooperate with each other instead of competing. It is different from the other types of equilibria in that it does not rely on the assumption that players are completely rational and make optimal decisions based on their individual payoffs. Instead, this type of equilibrium assumes that players will use a mixture of cooperation and defection, depending on the situation they find themselves in.

In comparison, a mixed-strategy equilibrium is an equilibrium in which players employ a combination of strategies instead of only one strategy in order to maximize their expected payoff. Mathematically, this can be represented as $P_i(s_i, s_{-i}) = \sum_{s_j} p(s_j) \cdot u_i(s_i, s_j)$ for all players i and all strategies s, where P_i is the expected payoff for player i, p is the probability distribution over the strategies employed by all players, and u is the utility function for player i. In contrast, bourgeois equilibrium does not require any probabilistic elements; rather it simply requires that each player select a single strategy that they believe will yield the best outcomes.

Correlated equilibrium (CE) is an extension of NE where each player's strategy depends on an additional set of random variables called "signals." Mathematically, this can be represented as $\sum_{a \in A} P(a) \cdot v(a|x) = v(x)$, where A is the set of possible action profiles, P is a probability distribution over those profiles, and v is the utility function for player i. CE is different from BE in that it allows for the possibility of coordination amongst the players, such as by having one player's strategy depend on another's. This coordination does not occur in bourgeois equilibrium, which instead focuses on each individual's strategy being independent from one another.

Evolutionary stable correlation (ESC) describes a situation in which two or more agents have adapted to cooperate with each other to achieve higher payoffs than either could achieve alone. Mathematically, this can be represented as $\max_{p \in \Delta}[U(p)]$, where Δ is the set of probability distributions over actions taken by agents and U represents their joint utility function.

Mathematically, BE is represented by a NE, which is defined as:

$$(s_1^*, s_2^*, ... s_n^*) \in S_1 \times S_2 \times ... S_n$$

where s_i^* represents the optimal strategy for player i. In contrast, a mixed-strategy equilibrium can be represented as:

$$(p_1, p_2, ... p_n) \in D$$

where D is the set of probability distributions over $S_1 \times S_2 \times ... S_n$. A correlated equilibrium (CE) can be represented as:

$$(s_{c1}, s_{c2}, ... s_{cn}) \in S_{C1} \times S_{C2} \times ... S_{Cn}$$

where S_{Ci} represents the set of strategies available to Player i given the coordination between players. And an evolutionary stable correlation (ESC) can be represented by a Nash Equilibrium with strictly dominant strategies:

$$(s^{**}, s^{**}, ... s^{**}) \in S'$$

where S' is the set of strict dominant strategies.

Interpretation of HDB: correlation

There are two main interpretations of bourgeois equilibrium: with correlation and with uncorrelated asymmetry. Let us look closer at each.

What is meant by "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 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 correlated equilibrium 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.

Correlated equilibria can be defined by the following equation:

$$\max_{x_1, ... x_n} \sum_{i=1}^{n} u_i(x_i)$$

$$\text{s.t.} \quad x_1 = c(y_1, ... y_{n-1}) \quad \forall i > 1 : x_i = c(y_i)$$

where u_i represents the utility function for each player i, x represents the strategy chosen by each player i, and y represents the common

randomizer chosen by all players. 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 correlated equilibria allow people to take into account each other's behavior when making decisions.

If bourgeois strategy is an ESS, it does not presuppose any randomization. However, many scholars studying the emergence of conventions interpret them as CE. Some researchers base their explanations on interpretation of HDB. For example, Guala (2016) defines social institutions as CE with normative force rooted in HDB. Gintis explicitly refers to HDB as not to strict NE, but as to CE (Gintis, 2009).

The intrinsic problem with bourgeois as CE 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 "bourgeois" strategy in the form "if own, then Hawk" and "if do not own, then Dove" is genuinely correlative. "Bourgeois" is correlated equilibrium 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).

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.

Evolution, Bayesian updating and correlation

Conclusion

References

Albert, R., & Thakar, J. (2014). Boolean modeling: A logic-based dynamic approach for understanding signaling and regulatory networks and for making useful predictions. *WIREs Systems Biology and Medicine*, *6*(5), 353–369. https://doi.org/10.1002/wsbm.1273

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

Baraghith, K. (2019). Emergence of Public Meaning from a Teleosemantic and Game Theoretical Perspective. *Journal of Philosophy*, 30.

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

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

Franke, M., & Wagner, E. O. (2014). Game Theory and the Evolution of Meaning: Game Theory and the Evolution of Meaning. *Language and Linguistics Compass*, *8*(9), 359–372. https://doi.org/10.1111/lnc3.12086

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

//books.google.com?id=yYvcDwAAQBAJ

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

Godfrey-Smith, P. (2014). Sender-Receiver Systems within and between Organisms. *Philosophy of Science*, *81*(5), 866–878. https://doi.org/10.1086/677686

Guala, F. (2016). *Understanding institutions: The science and philosophy of living together*. Princeton University Press.

Harms, W. F. (2004). *Information and Meaning in Evolutionary Processes*. Cambridge University Press. https://books.google.com?id=ztl99e9ugtAC

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

Huttegger, S. M. (2007). Evolutionary Explanations of Indicatives and Imperatives. *Erkenntnis*, *66*(3), 409–436. https://doi.org/10.1007/s10670-006-9022-1

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. (2008). *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

Millikan, R. G. (1987). *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press. https://books.google.com?id=jncHBAYe8TkC

Millikan, R. G. (2005). *Language: A biological model*. Clarendon Press ; Oxford University Press.

Neander, K. (2008). Teleological Theories of Mental Content: Can Darwin Solve the Problem of Intentionality? In M. Ruse (Ed.), *The Oxford Handbook of Philosophy of Biology* (p. 0). Oxford University

Press. https://doi.org/10.1093/oxfordhb/9780195182057.003.0017

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.

O'Connor, C. (2021). Measuring Conventionality. *Australasian Journal of Philosophy*, *99*(3), 579–596. https://doi.org/10.1080/00048402.2020.1781220

Planer, R. J., & Godfrey-Smith, P. (2021). Communication and representation understood as sender–receiver coordination. *Mind & Language*, *36*(5), 750–770. https://doi.org/10.1111/mila.12293

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

Seyfarth, R., & Cheney, D. (1990). The assessment by vervet monkeys of their own and another species' alarm calls. *Animal Behaviour*, *40*(4), 754–764. https://doi.org/10.1016/S0003-3472(05)80704-3

Shea, N. (2018). *Representation in cognitive science* (First edition). Oxford University Press.

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. (2010). *Signals: Evolution, learning, & information*. Oxford University Press.

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

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

Sterelny, K. (2017). From code to speaker meaning. *Biology & Philosophy*, *32*(6), 819–838. https://doi.org/10.1007/s10539-017-9597-8

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

Zawidzki, T. W. (2013). *Mindshaping: A New Framework for Understanding Human Social Cognition*. The MIT Press. https://doi.org/10.7551/mitpress/8441.001.0001