Evolution of social institutions from “animal conventions” with niche construction in Guala’s unified social ontology

phd thesis work-in-progress

$date

Future social science may be as ontologically surprising as twentieth-century physics was

Don Ross

He who understands baboon would do more towards metaphysics than Locke

Charles Darwin

# Introduction

## Relevance of the study

Both animals and humans coordinate their behavior strategically. At least, it can be described as such by game theory. However, some game-theoretic models of human and animal coordination are formally indistinguishable. For example, consider two butterflies competing over the mating spot, where one is the “owner” and another is “intruder” of the spot. How do they resolve this dispute? As research shows, they mostly take into account who occupied the spot first (Davies 1978). This makes each butterfly capable of following one of the *conditional strategies*:

* “if I am an owner of this spot, I will fight over it whatever it takes”
* “if I am an intruder, I will show hostility but will not escalate and eventually retreat”.

Now consider a similar situation in humans. Two tribes graze their cattle on different sides of a dried riverbed. When there was a river, they both *knew* that there is another tribe over the river and it is in their best interest not to trespass. After the rived has dried, they kept using a physical marker (dried riverbed) to construct a more abstract conditional strategy “if the land is to the South (or North for another tribe) from the riverbed, graze the cattle” (Guala and Hindriks 2015). Thus, they resolve a potential conflict around their property in a similar way by following a conditional strategy. Structurally and formally, these cases look the same:

$$
\begin{array}{|c|c|c|c|}
\hline
& Fight & Retreat & Conditional \\
\hline
Fight & 0,0 & 2,1 & 1,0.5 \\
\hline
Retreat & 1,2 & 1,1 & 1,1.5 \\
\hline
Conditional & 0.5,1 & 1.5,1 & 1.5,1.5 \\
\hline
\end{array}
$$

Miscoordination game with conditional strategies

In both cases, the agents have conditional strategies, but what is hidden is the nature of *conditionalization*. In the first case, it relies solely on environmental features and in the second, it relies on explicit shared normative beliefs. Butterflies coordinate on order of precedence and do not “think” about it, while we humans not only think (sometimes) but base the very game and strategies on our *representations* of others’ actions and beliefs. So, the difference is in *on what agents condition their strategies*.

The thing on which agents condition their strategies is called *correlation device* in game theory. And in this dissertation, I investigate the cognitive evolution of correlation devices: how agents, throughout the evolution, learn to coordinate on increasingly complex correlation devices — from simpler environmental cues like “occupied spot” or “opponent’s size” to shared normative beliefs represented as rules like “North to the river is outher tribe’s property”.

Although it might look like the tribe members condition their actions purely on a physical marker of a dried riverbed just like animals would do, it is not the case. A physical marker such as a dried riverbed, as I will argue, *scaffolds* a deeper epistemic structure involving *common reason to believe* which land belongs to whom. It *encodes* information so that agents coordinate not on the riverbed itself but the associated normative rules. Therefore, it is a correlation device *abstracted* from a physical signal like “opponent’s size” or even “dried riverbed”.

From a formal point of view, the problem seems almost invisible: regardless of the complexity of correlation device, whether it is “opponent’s size” or “shared beliefs about each other’s property”, the payoffs of the game stay the same. What is changing, though, is *informational structure of environment* in which the game is played which affect the nature of correlation devices used in the game, either simpler and “flat” or more complex and abstracted.

As I will argue after Sterelny (2003) and Godfrey-Smith (Planer and Godfrey‐Smith 2021; Godfrey-Smith, n.d.), it is increasing informational complexity of environment that drives adaptive pressure on simple non-epistemic agents to evolve cognitive capacities for representation of more abstract features of environment to coordinate upon, thus driving the evolution of more complex correlation devices like social norms from simple environmental cues.

Most importantly, I will argue that such co-evolution of cognitive capacities and informational structure of environment drives evolution of human social institutions as belief-mediated norm-governed behavioral patterns from animal “conventions”, connecting naturalistic social ontology (Guala 2016) and philosophy of biology (Sterelny 2003; Skyrms 1994).

Such a project does three things, unfolding progressively:

1. In social ontology, it resolves an inner tension within naturalistic social ontology, which tries to explain the social stability by appealing to equilibria in strategic games (Guala 2016; Frank Hindriks and Guala 2015). The tension lies in the conflation of underlying assumptions about the properties of agents playing a game from different strands of game theory;
2. In social ontology, it defends a view on social institutions as *natural kinds* proper[[1]](#footnote-20). Although there have been proof-of-concept accounts modeling evolution of social institutions (Skyrms 2003, 2014a; Bowles, Choi, and Hopfensitz 2003), they do not incorporate *cognitive realism* and actual mechanism of cognitive evolution as prerequisites for emergence of uniquely human social institutions as scalable norm-governed practices. We explicitly discuss and model *cognitive* evolution of social institutions;
3. In philosophy of social science, showing how social institutions might have evolved in a cognitively realistic way can provide a naturalistic ground for sociological theory building and, in particular, for ontological constraints on it, contributing to analytically core concepts independent from epistemological optics akin to natural sciences.

More generally, the relevance of the work stems from its explicit modeling of *emergence of social epistemology*, which is largely said to ground social ontology (J. Searle 1995; Guala 2016; Gintis 2009b; Tuomela 2013) and to be one of the most important questions on philosophy of social science.

Specifically, my question in the current thesis is *how the cognitive architecture underlying normative behavior which requires sophisticated representational capacities could have evolved from simpler representational systems and forms of social coordination*?

The project is situated at the intersection of social ontology, evolutionary game theory and philosophy of biology. It seeks to contribute to the ongoing endeavor to naturalize social normativity, that is, to explain normative phenomena as arising from natural processes without metaphysical or purely interpretive notions. This aligns with a scientific realist perspective that treats social kinds, such as institutions and norms, as real entities with causal efficacy, amenable to empirical investigation and theoretical modeling (**boyd1999?**; Guala 2016).

The dissertation builds on Francesco Guala’s *rules-in-equilibrium* theory (RiE), which provides a unified framework for understanding social institutions as *correlated equilibria* (CE)[[2]](#footnote-21) maintained by agents’ conditional strategies (Guala and Hindriks 2015). However, Guala’s theory, while conceptually strong, leaves open the question of how such equilibria and the underlying conditional strategies evolve from simpler forms of coordination observed in non-human animals when the payoff matrices are identical as we saw earlier. Addressing this gap requires integrating insights from evolutionary game theory and philosophy of biology concerning cognitive evolution to trace the evolutionary emergence of epistemic belief-mediated agency.

A central contribution of this dissertation is a naturalistic and evolutionary account of how humans come to represent and follow institutional rules as prescriptions, thereby addressing a key lacuna in Guala’s RiE theory of institutions (Guala 2016). His account unifies the “institutions-as-rules” (J. Searle 1995) and “institutions-as-equilibria” (North 1990; Aoki 2007) traditions, arguing that institutions are best understood as rules that are in equilibrium: they are both cognitive devices (rules) that individuals use to represent and navigate social life, and stable patterns of behavior (equilibria) that persist due to mutual expectations and incentives. While this approach elegantly bridges philosophical and game-theoretic perspectives, Guala’s theory leaves underexplored the cognitive and evolutionary mechanisms by which agents come to *internalize equilibria as rules*, and how the prescriptive force of rules arises from mere behavioral regularities.

The current dissertation advances Guala’s project by explicitly modeling the evolutionary pathway from coordinating on simple physical cues to the emergence of decoupled, belief-like representations of rules that agents act upon. Drawing on recent developments in game theory (Skyrms 1994, 2010a) and the philosophy of cognitive science (Sterelny 2003, 2012b, 2021; Godfrey-Smith 2002; Planer and Godfrey‐Smith 2021; Seitz 2020), the thesis shows how agents, starting as simple reinforcement learners, evolve to infer the latent structure of their environment, develop *decoupled representations*, and ultimately coordinate on shared rules that are experienced as normative. This approach substantiates Guala’s claim that rules are cognitive representations of equilibria, but, crucially, it demonstrates *how such representations can evolve from non-normative origins*, rather than being presupposed or imposed by external correlation devices.

Furthermore, by formalizing the emergence of belief-mediated coordination within agent-based and evolutionary game-theoretic models, the thesis addresses critiques that Guala’s theory is overly externalist (**roversi2021?**) or insufficiently accounts for the material and psychological reality of institutions (Rabinowicz 2018). It provides a detailed account of how rule-like character of institutions can be grounded in evolved cognitive capacities for decoupled representation, thus offering a more robust naturalistic ontology of institutions that integrates both their behavioral and normative dimensions.

Beginning with the foundational models of coordination by David Lewis (Lewis [1969] 2008), Brian Skyrms (Skyrms 1994, 2010a) and others, and extending with scrutiny of Guala’s game-theoretic social ontology, the analysis leads to introduce a crucial distinction between *ontic correlation of strategies* — ecological, cue-driven mechanisms of coordination leading to *CE*, and *epistemic correlation* — cognitive, representation-based processes that enable agents to interpret and respond to social cues which lead to epistemic CE. Formally, these are the same, as agents still play stable and optimal conditional strategies, but what is “hidden” is the nature of correlation device which signals each player a preferable strategy.

The main novelty (which is also a difficulty for comprehending it) of this dissertation lies in providing *causal mechanisms of evolution of complex belief-based correlation devices* which lie *“behind”* the game structure: the agents, their strategies and payoffs stay the same throughout the evolution. What changes is *the nature of the correlation device* on which the agents coordinate and follow conditional strategies.

Formally, both animals and humans use some “signal”, or environmental asymmetry, be it “opponent’s size” for the former and traffic lights for the latter. But the underlying *epistemology of these signals* in both cases is what matters most in this work.

The concept of *social convention* plays a major role in this dissertation and is understood in line with the tradition of David Hume and David Lewis as a pattern of behavior that arises within a population through repeated interaction, facilitating coordination and reducing uncertainty. *Normativity* refers to the prescriptive force that such conventions can acquire over time, often accompanied by expectations, sanctions, or moral justification. The distinction between the two is not always clear-cut, and one of the aims of this dissertation is to clarify the mechanisms by which conventions become normative.

The *relevance* of this research in the contemporary context is multifaceted. Theoretically, the results may be of interest to scholars in philosophy of science, social ontology, cognitive science, and evolutionary theory. Practically, the insights gained may inform the design of institutions (by accounting for their emergent cognitive dynamics), the development of artificial intelligence systems, and the analysis of organizational dynamics. Pedagogically, the results of the research can be used as materials for courses on philosophy of social science and social ontology.

## Object and subject of study

The *object of this research* is Guala’s *rules-in-equilibrium* theory, which offers a unified social ontology that synthesizes two influential traditions in the philosophy of social institutions: the institutions-as-rules approach and the institutions-as-equilibria approach (Guala and Hindriks 2015). This theory conceptualizes social institutions as sets of regulative rules which are conditional strategies of the form “if X do Y” which coordinate behavior by creating CE in strategic interactions. Unlike earlier models that treat institutions either as mere conventions or as constitutive rules, Guala’s framework emphasizes the causal role of CE sustained by agents’ conditional strategies as *beliefs*. This approach is particularly valuable because it provides a rigorous and philosophically nuanced account of how social order can emerge and persist without presupposing idealized rationality or collective intentionality as a lot of contemporary social ontology does (Jankovic and Ludwig 2018).

Building on Guala’s own attempt to naturalistically ground his theory, this dissertation takes seriously the question of the *naturalistic underpinnings* of Guala’s framework. More specifically, it investigates how social institutions, as described by rules-in-equilibrium theory, could have evolved from animal conventions within an evolutionary and ecological context. While Guala’s theory is primarily conceptual and slightly formal, it leaves open the empirical and mechanistic question of how such institutions arise in real populations of agents with limited cognitive capacities. The present research aims to fill this gap by exploring the evolutionary trajectory from simpler forms of correlated coordination observed in and modeled from non-human animals to the complex normative institutions characteristic of human societies. This involves examining whether and how the CE that sustain institutions can be grounded in selection processes, ecological constraints, and cognitive evolution.

The *subject of the study* is the *evolutionary emergence of epistemic belief-mediated conditional strategies* from more primitive, reactive, non-epistemic conditional strategies within Guala’s rules-in-equilibrium theory. In particular, its emergence within the game-theoretic settings of the Hawk-Dove-Bourgeois game which I mentioned in the beginning. Reactive conditional strategies are those that depend directly on environmental cues without involving internal representations or beliefs about others’ intentions or future actions. For example, a butterfly conditioning its behavior on the immediate presence of a signal or a stimulus like “is there anyone already in this spot?”. It is direct perception of a cue and not its cognitive representaitonal processing. In contrast, epistemic belief-mediated conditional strategies involve agents forming internal models or beliefs about the behavior of others, allowing for more flexible, anticipatory, and context-sensitive responses. Tribes grazing their cattle form their strategies according to shared beliefs about each other’s property abstracted from immediate observable borders.

This transition from reactive to epistemic strategies marks a crucial evolutionary step toward what can be called *proto-normativity* which is a stage where agents’ conditional strategies are not merely responses to stimuli but are mediated by deontic or proto-deontic beliefs about what ought to be done in a given context. This proto-normativity does not yet encompass fully developed normative systems with explicit mechanisms for sanctioning deviations or enforcing compliance but represents the foundational cognitive and behavioral capacities that make such systems possible. By focusing on this emergence, the dissertation contributes to understanding how the cognitive architecture underlying normative behavior could have evolved gradually from simpler forms of social coordination.

## Goals and tasks of the study

The overarching *goal* of this dissertation is to provide a *naturalistic account of the emergence of epistemic belief-mediated agency and proto-normativity*, thereby grounding the possibility of “proper” social normativity within an evolutionary framework. This goal is motivated by the recognition that many existing accounts of normativity either presuppose advanced cognitive capacities such as the ability to represent collective intentionality or to enforce sanctions or treat norms as purely social constructs without sufficient grounding in natural processes somehow already existing in the environment.

This research aims to bridge that gap by focusing on the *initial evolutionary and ecological processes* that give rise to proto-deontic beliefs about what one ought to do-that underpin normative behavior. These proto-normative beliefs are understood as internal states that mediate conditional strategies, enabling agents to coordinate behavior not reactively but through anticipatory and belief-based mechanisms.

Unlike accounts that emphasize punishment or enforcement as central to normativity, this dissertation concentrates on the *emergence of the cognitive and behavioral substrates* that make such systems possible. By doing so, it lays the groundwork for future research on how proto-normativity can evolve into fully-fledged normativity with social enforcement mechanisms, understood naturalistically.

More concretely, the study seeks to:

* Scrutinize the relationship between social conventions and normativity to better articulate distinctions within Guala’s theory
* Analyze Guala’s theory from epistemological point of view to spot inconsistencies which hide the question of different cognitive architectures inside his theory
* Delineate different cognitive architectures presupposed by Guala’s theory and elucidate plausible evolutionary pathways for connecting them
* Demonstrate, with agent-based simulation, how epistemic correlation can emerge from ecological interactions and selective pressures favoring more sophisticated representational capacities
* Clarify the conceptual relationship between reactive coordination mechanisms and epistemic, belief-mediated strategies
* Situate these findings within Guala’s rules-in-equilibrium framework, thereby providing a naturalistic foundation for the institutions it describes
* Relate findings with debates about metaphysics of social kinds and philosophy of social science.

## Main research hypotheses

The central *hypothesis* of this dissertation can be stated as follows:

*Epistemic belief-mediated conditional strategies that depend on agents’ internal beliefs about others’ behavior can evolve from reactive, non-epistemic conditional strategies based on environmental asymmetries through ecological and evolutionary processes of cognitive niche construction that can be rigorously modeled within game-theoretic frameworks with agent-based modeling. This evolutionary transition underpins the emergence of proto-normativity, providing a naturalistic basis for social institutions conceptualized as rules-in-equilibrium.*

This hypothesis entails several interrelated claims:

1. *Reactive conditional strategies* that respond directly to environmental cues without involving internal representations constitute the evolutionary antecedents of more complex, belief-mediated strategies. Such reactive strategies are widespread in non-human animals and form the baseline for social coordination.
2. Through *repeated interactions and selective pressures* favoring successful coordination and cooperation, agents develop *internal epistemic states* like beliefs or representations that mediate their conditional strategy choice. These epistemic states enable agents to predict and respond flexibly to the behavior of others, rather than merely reacting to fixed cues.
3. The emergence of these epistemic strategies gives rise to *proto-normative attitudes* towards what actions are appropriate or required in given social contexts. These proto-normative beliefs are the cognitive precursors to full-fledged normative systems.
4. The entire process can be *modeled using agent-based simulation*, demonstrating the plausibility of emergence of epistemic correlation devices without presupposing more advanced cognitive constructs like *collective intentionality* or explicit sanctioning mechanisms.
5. This transition is not merely conceptual but can be *computationally investigated* through agent-based simulations that capture the dynamics of the cognitive evolution of correlation devices represented as ecological cues of different layers of abstraction.

## Theoretical basis of the research

This dissertation draws on a rich and interdisciplinary theoretical foundation that spans social ontology, philosophy of science, evolutionary biology and cognitive science. Key components include:

* *Guala’s rules-in-equilibrium theory:* Guala’s framework offers a philosophically grounded account of social institutions as CE sustained by agents’ conditional strategies (Guala and Hindriks 2015). By integrating the normative and causal dimensions of institutions, this theory strives to provide a parsimonious naturalistic ontology of social order that avoids taking metaphysically-laden constitutive rules or collective intentionality as primitives. The present research builds on this foundation by exploring the evolutionary origins of the conditional strategies that sustain these equilibria.
* *Evolutionary game theory:* The formal modeling of strategic interactions and their evolution over time is central to this study. Foundational work by Lewis ([1969] 2008), Maynard Smith (1982), Skyrms (2010a) and Gintis (2009a) has demonstrated how cooperation and coordination can emerge through evolutionary processes. This dissertation extends these insights by focusing specifically on the evolution of epistemic, belief-mediated strategies from reactive ones, using CE as an analytical tool.
* *Philosophy of biology*: Understanding the evolution of internal representations and belief systems is critical for distinguishing reactive from epistemic strategies. Sterelny (2003, 2012b, 2021) has emphasized the role of *decoupled representation* in cognitive evolution of humans. Millikan’s teleosemantic theory (**millikan1984?**) provides a naturalistic account of mental representation grounded in biological function which serves a foundation for Sterelny’s theory. Another important source is Godfrey-Smith’s *Environmental Complexity Thesis* (Godfrey-Smith, n.d.; Godfrey-Smith, Sternberg, and Kaufman, n.d.), which states that the very function of cognition is navigating in informationally rich environments and producing adaptive repose enhancing one’s own fitness. These perspectives inform the conceptualization of epistemic conditional strategies as belief-mediated in the current thesis.
* *Social ontology and philosophy of social science:* The metaphysical and normative implications of treating social conventions as natural kinds are informed by the work of Boyd (**boyd1999?**), Searle (1995), and contemporary debates on the naturalization of normativity. Guala’s RiE provides a reference point for situating the present research within ongoing philosophical discussions being “a central hub” and synthesizing major findings across the subdisciplines of philosophy of science.

Together, these theoretical strands enable a rigorous, interdisciplinary investigation into the naturalistic emergence of epistemic correlation devices as proto-normative beliefs, bridging gaps between animal coordination and human social institutions within Guala’s unified social ontology.

## Prior research on the topic

Early foundational work by Lewis ([1969] 2008) introduced game-theoretic formalization of social conventions as stable mutually beneficial behavioral patterns, emphasizing how coordination equilibria arise through repeated interactions. Skyrms expanded this framework by incorporating evolutionary dynamics, demonstrating how populations of agents can evolve stable conventions through selection processes leading to beneficial social contract arrangements Skyrms (2010a). These models successfully explain the emergence of coordination but often get criticized for excluding normativity of conventions from its explanatory scope and relying on cognitively implausible agents with idealized rationality assumptions (Sterelny 2012c; Gilbert 1992; Hédoin 2021). The detailed inquiry into the historical approaches to social conventions and their normativity in formal philosophy will be presented in [Chapter 1](#X5dfcd333cac83d7048e2b6598f0f29592038c69) dedicated fully the Humean/Lewisian tradition of game-theoretic treatment of social conventions.

Guala’s *rules-in-equilibrium* theory (Guala and Hindriks 2015) represents a significant advance by integrating the normative and causal dimensions of social conventions with classic social ontology of Searle (J. Searle 1995). As mentioned before, while providing a powerful intergartive framework for social ontology, Guala’s theory does not explicate the evolutionary and cognitive mechanisms by which conditional strategies arise. This gap has motivated subsequent research exploring the naturalistic foundations of social institutions (Hédoin 2021). We will analyze Guala’s theory in detail in Chapter 2, surfacing the problems of conflating agency models implied in his arguments and of conflating the notions of representation he uses to delineate animal and human social coordination.

Parallel developments in cognitive science and philosophy of biology have highlighted the importance of epistemic capacities, such as decoupled representation. These perspectives emphasize that proto-normativity involves not only behavioral coordination but also internal states-beliefs and intentions that mediate social interaction (Sterelny 2021). Yet, the evolutionary transition from reactive, cue-driven behaviors to epistemic, belief-mediated strategies remains insufficiently understood. We will address theories of cognitive evolution (Sterelny 2003, 2012b; Godfrey-Smith, n.d.; Ruth Garrett Millikan 1987) and cognitive niche construction (Odling-Smee, Lala, and Feldman 2003; Bardone and Magnani 2007; Planer and Godfrey‐Smith 2021) in Chapter 3, where we review them in relation to the Guala’s initial attempt at using Sterelny’s arguments for decoupled representation to ground the difference between animal and human coordination (Frank Hindriks and Guala 2015). There, we will also look at existing game-theoretic attempts at incorporating niche construction into its models to explain the stability of social institutions and identify their inefficiencies for our account. Afterwards, we provide our own computational model of cognitive evolution of epistemic belief-mediated correlation of strategies with niche construction.

Overall, while the literature provides valuable insights and diverse approaches, a comprehensive, mechanistic account that integrates evolutionary game theory, philosophy of biology and social ontology remains underdeveloped. This dissertation aims to address this lacuna by offering a detailed, naturalistic model of the evolutionary emergence of epistemic belief-mediated conditional strategies, thereby advancing the understanding of proto-normativity and its role in social institutions.

## Methodology

The methodology adopted in this research is interdisciplinary, combining conceptual analysis, formal modeling, and computational simulation to investigate the emergence of proto-normativity within an evolutionary framework.

1. *Conceptual analysis:* The dissertation begins by clarifying key concepts which involves critical engagement with existing philosophical literature to establish a coherent conceptual framework that guides the subsequent work.
2. *Comparative theoretical analysis:* existing models of social norm emergence, including Lewis’s signaling games, Skyrms’s evolutionary models and accounts of Ullmann-Margalit (1977), Young (1998), J. M. Epstein and Axtell (1996), Bicchieri (2005), Gintis (2009b) and Guala’s rules-in-equilibrium theory (Guala 2016) are systematically compared and critiqued. This comparative approach identifies strengths, limitations, and gaps that the present research seeks to address.
3. Agent-based simulation\*: building on the foundational work in evolutionary game theory (Maynard Smith 1982; Skyrms 1994), the dissertation develops a formal model that simulates populations of agents employing conditional strategies and “growing” epistemic conditionalization as an adaptive response to increasing informational density of environment. This model explores how reactive strategies can evolve into belief-mediated ones under ecological pressures. Correlated equilibrium, evolutionary stable strategies and Hawk-Dove-Bourgeois game serve as analytical tools for understanding the stability and causal efficacy of emergent strategies.

## Novelty of the research

This dissertation offers several novel contributions to the study of social normativity, social ontology and philosophy of social science:

* *Ontic/epistemic correlation:* the introduction and rigorous development of the *ontic/epistemic correlation* distinction as a key explanatory tool for understanding the evolutionary transition from reactive to belief-mediated conditional strategies is original. This distinction clarifies the cognitive and ecological mechanisms like niche construction underlying the emergence of advanced cognitive architectures capable of proto-normativity.
* *Mechanistic transition from ontic to epistemic correlation*: the research provides a mechanistic account of how epistemic conditional strategies could have evolved naturally from simpler reactive behaviors, filling a gap in existing literature.
* *Naturalistic grounding of Guala’s theory:* the dissertation extends Guala’s *rules-in-equilibrium* framework by explicating its naturalistic foundations, thereby bridging unified social ontology with evolutionary and cognitive science.
* *Support for institutions as natural kinds*: the dissertation supports the thesis that social institutions are mind-independent natural kinds, which do incorporated epistemic element affecting its shape and function. By providing a mechanistic model of emergence of epistemic correlation devices from ontic ones, we show how institutions can be both grounded in natural evolutionary processes and have a mind-dependent interactive component.

Thorough literature reviews in respective chapters confirm that no prior work has combined these conceptual, formal, and empirical elements in a unified investigation of the evolutionary emergence of epistemic belief-mediated conditional strategies within a naturalistic social ontology.

## Theses for defense

The following theses summarize the core claims of this dissertation and are proposed for defense:

1. Social institutions, as conceptualized by Guala’s *rules-in-equilibrium* theory, can be grounded in evolutionary processes aided with cognitive niche construction that produce normative correlated equilibria sustained by epistemic agents’ conditional strategies
2. The ontic/epistemic correlation distinction is essential for explaining the evolutionary transition from ontic to epistemic conditional strategies in Guala’s unified social ontology
3. Agents with a capacity for information integration via functional abstraction are more evolutionarily robust than other agents
4. Social conventions and proto-normative beliefs qualify as natural kinds within a scientific realist social ontology
5. The naturalization of social ontology is possible without collective intentionality or explicit sanctioning mechanisms and just with ecological dynamics: co-evolution of cognitive capacities for information integration and environmental change induced by agents.

## Conclusion and acknowledgments

The dissertation undertakes a modest yet rigorous investigation into the naturalistic emergence of cognitive architecture capable of proto-normativity and its role in grounding social institutions within Guala’s unified social ontology. By developing a conceptual distinction between ontic and epistemic correlations and employing formal evolutionary game-theoretic models alongside agent-based simulations, the research offers a mechanistic account of how epistemic belief-mediated conditional strategies evolve from reactive behaviors.

Ultimately, this work aspires to contribute to a more integrated and naturalistic understanding of social order, bridging gaps between animal coordination and human normativity, and fostering dialogue across disciplinary boundaries.

I thank:

* **my advisor Vitaly Dolgorukov** for urging me to be more rigorous and pay attention to formal aspects of the work;
* **my first advisor Helena Knyazeva** for encouragement and support for exploring this inherently complex topic during my early PhD days;
* **Francesco Guala** for insightful discussions and comments on the parts of the current work;
* **Brian Skyrms** for important clarifications about the relationship between evolutionary stable strategies and correlated equilibria in his work (which caused a lot of confusion for me);
* **Don Ross** for discussions of the relationship between social coordination and cognition;
* **Christophe Heintz** for deep questions and genuine interest in my work;
* **Stephen Turner** for careful and lengthy comments on my early drafts and discussions about coordination, cogntition and philosophy of social science;
* **colleagues at conferences** who asked questions and engaged in the discussions afterwards.

I also have infinite gratitude to my partner and family for support and care during my PhD years. Thank you!

# **Chapter 1.** Social conventions: Hume, Lewis and game theory

The tradition of understanding social coordination as a source of social order is historically rich. Aristotle grounded social conventions in human nature and the pursuit of *eudaimonia*, or flourishing. He viewed humans as “political animals” who naturally form communities to achieve collective well-being. Justice and virtue, central to his ethics, were seen as the basis for political order. Unlike later followers of the social contract theory, Aristotle saw social organization as intrinsic to human rationality rather than a deliberate agreement (**aristotle1998?**).

Hobbes reimagined social conventions as constructs invoked by humanity’s violent “state of nature.” He argued that self-preservation drives individuals to surrender freedoms to an absolute sovereign via a social contract resulting from explicit *agreement* (Hobbes [1650] 2016). Conventions thus arise from fear and rational self-interest, not innate sociability.

According to B. Epstein (2018), a notion of *convention* was first explicitly used as an alternative to agreement by Pufendorf (1673), to refer to language and law. He synthesized Hobbesian ideas with theological natural law. While agreeing that humans are self-interested, he attributed the “law of sociality” to divine mandate, requiring peaceful coexistence despite innate corruption. For Pufendorf, natural law obligates humans to form civil societies, with God as the ultimate author of social conventions. This introduced a moral dimension absent in Hobbes’s instrumentalist framework, suggesting that conventions are not merely utilitarian but also morally justified. His point was that conventions do not need to be explicitly agreed to and might exist and work without their intentional design. This intuition has remained largely unchanged.

Hume’s theory of social conventions (Hume 2003, 1998) offers a groundbreaking empiricist account of how conventions emerge organically from human interaction rather than rational design or divine mandate. Hume’s analysis hinges on three core premises:

* the role of custom in shaping behavior
* the centrality of mutual benefit
* the artificiality of conventions.

These are seen as products of collective habit rather than explicit verbal agreement. The components form the scaffolding of his theory, which bridges psychology, ethics, and political philosophy.

Hume’s empiricist framework posits that human understanding arises from sensory impressions and ideas derived from them. This extends to social behavior: conventions emerge not from reason but from repeated experiences that cultivate habits. For instance, Hume’s iconic example of two individuals rowing a boat illustrates how synchronization arises through trial and error, not prior negotiation:

“Two men who pull at the oars of a boat, do it by an agreement or convention, tho’ they have never given promises to each other”(Hume 2003).

However, Schliesser (2024) stipulates that this kind of coordination is not backed by “Humean conventions”, for they, according to Hume himself[[3]](#footnote-33), require “positive social externality”, whereas two burglars could effectively row away from a crime scene. We will not focus on this morally-driven notion of conventions.

Over time, repeating patterns solidify into conventions because they resolve practical problems like coordinating labor and establishing property rights while minimizing friction. Custom, as Hume writes, “renders our experience useful to us” by creating stable expectations about others’ behavior, even in the absence of formal rules (Hume 2003) . This emphasis on habit challenges rationalist theories like Hobbes’s by showing how conventions evolve *unconsciously* through iterative adjustments.

Hume highlights four key features of conventions:

* *Mutual benefit*: all parties gain from adhering to the convention (e.g., synchronized rowing ensures progress; standardized currency facilitates trade)
* *Multiple potential solutions*: different solutions could theoretically work (e.g., rowing fast or slow), but consistency matters more than specific choice
* *Unplanned agreement*: conventions develop spontaneously through “a slow progression” of trial and error, not deliberate contract
* *Reciprocity*: adherence to to convention depends on the expectation that others will reciprocate, creating a self-reinforcing cycle of trust.

For Hume, conventions like property rights arise because humans recognize the “common interest” in stabilizing possessions to avoid conflict, even if their natural inclinations lean toward self-interest (Hume 1998). This pragmatic focus distinguishes his theory from moralistic accounts, framing conventions as tools for managing inherent human partiality.

Hume classifies conventions as *artificial virtues*, social constructs developed to counteract humanity’s “limited generosity”. Unlike natural virtues like benevolence, which arise instinctively, conventions like justice or promise-keeping require cultivation. Their artificiality, however, does not make them arbitrary. Instead, they gain normative force through collective sentiment: individuals approve of conventions that promote social utility, and disapproval of violations strengthens adherence over time. This process explains how conventions acquire moral weight, transforming into norms that feel binding even when rational self-interest might suggest defiance. Experimental studies inspired by Hume’s (or rather Lewis’s (**lewis2008?**)) work confirm that conventions stabilize behavior even when incentives to defect arise, underscoring the interplay of habit and normativity (Guala and Mittone 2010).

Hume’s theory diverges sharply from social contract models. While Hobbes rooted conventions in deliberate agreements to escape chaos or secure rights, Hume dismissed the notion of a primordial “state of nature” requiring such pacts. Instead, he argued that conventions emerge incrementally from lived experience, reflecting his broader skepticism toward rationalist abstractions. His framework also anticipated modern game theory, particularly Lewis’s analysis of conventions as coordination equilibria (**lewis2008?**), though Hume placed greater emphasis on psychology.

Crucially, Hume’s account bridged descriptive and normative domains. By showing how conventions evolve from practical needs to moral norms, he offered a fairly naturalistic explanation for social order that avoids appeals to divine law or metaphysical necessity. This aligns with his rejection of causation as anything beyond observed regularity, reinforcing his view that human institutions are contingent products of custom rather than eternal truths.

After Hume, philosophers in the Scottish Enlightenment held that social order is an emergent product of individuals’ interactions, however, no such order has been specifically intended by individuals. As Ferguson ([1767] 1980) wrote, “nations stumble on establishments which are, indeed, the result of human action, but not the execution of any human design”. Afterwards, however, the study of conventions has quieten.

Lewis has revived and operationalized Hume’s insights into a theory of conventions using game theory and treating conventions as equilibria sustained by common knowledge and precedent. While Hume emphasized historical contingency and gradual emergence, Lewis imposed stricter criteria of rationality and mutual expectations (**lewis2008?**). He saw conventions as solutions to coordination problems, a class of problem in game theory (a branch of mathematics dealing with strategic behavior) which require two or more agents to align their actions to produce a jointly optimal outcome. In the next section, we will tour game theory and its main concepts before getting back to Lewis’s theory of conventions as game theory will be crucially important in the remainder of the thesis.

## Game theory background

Game theory is a mathematical framework used to analyze situations of strategic interaction between rational decision-makers. Originally developed by John von Neumann and Oskar Morgenstern in their seminal work *Theory of Games and Economic Behavior* (**morgenstern1944?**), game theory has since evolved to encompass a wide range of applications in economics, biology, political science, and sociology (Gintis 2009a; **osborne2004?**). It provides the tools to study how individuals or groups make choices when their outcomes depend not only on their own decisions but also on the decisions of others. The fundamental building blocks of game theory are games, players, strategies, payoffs, and equilibria (Zamir, Maschler, and Solan 2013).

A strategic game in game theory is defined as a formal model where:

* is a set of players
* is strategy sets of each player, where is the set of strategies available to player
* specifies the payoff functions, where gives the utility for player given the chosen strategy profile (**myerson1991?**).

A strategy is a complete plan of action a player will follow in any situation they might face within the game. Payoffs represent the rewards or utilities that players receive based on the combination of strategies chosen by all involved.

One of the central concepts in game theory is equilibrium, where no player has an incentive to unilaterally change their strategy given the strategies of others. The most well-known equilibrium concept is the Nash equilibrium (NE), introduced by John Nash in the early 1950s (**nash1950?**). A strategy profile forms a Nash equilibrium if for every player , the following condition holds:

Here,

* is a payoff function for player
* is a strategy chosen by player at equilibrium
* is a combination of strategies chosen by all other players except player

The inequality states that player cannot increase their payoff by unilaterally changing their strategy from to any other available strategy .

Shortly after Nash’s work, Robert Aumann introduced the concept of *correlated equilibrium* (CE) in 1974 (Aumann 1974). This generalization of Nash equilibrium allows players to coordinate their strategies through signals from a trusted mediator. Unlike Nash equilibrium, where players act independently, CE enables communication or correlation of strategies, capturing coordination through shared information. In a CE, a random signal suggests a strategy to each player, and players follow the recommendation if it is in their best interest to do so. Formally, a correlated equilibrium satisfies:

Here,

* represents the probability that the mediator recommends strategy to player and to the other players
* is the payoff to player when they play and the others play

The inequality ensures that the expected payoff from following the recommendation is at least as great as from deviating.

As Roger Myerson has reportedly observed,

“If there is intelligent life on other planets, in a majority of them, they would have discovered correlated equilibrium before Nash equilibrium” (Solan and Vohra, n.d.).

CE can be a more natural concept than Nash equilibrium, as its mathematical simplicity and reliance on cooperation make it easier to discover. Myerson argued that humanity’s prioritization of Nash equilibrium may have been an accident of history rather than a reflection of its fundamental importance. In societies or civilizations where cooperative behavior is emphasized or external mediators are prevalent, CE could emerge as a more intuitive starting point for understanding strategic interactions.

In the realm of evolutionary biology, John Maynard Smith introduced the concept of *evolutionarily stable strategy* (ESS) in 1973 (**maynard1973?**). An ESS is a strategy that is robust against invasion by mutant strategies and satisfies the following condition:

Here,

* is the payoff when both the incumbent and the invader use strategy .
* is the payoff when the invader uses strategy while the incumbent sticks to .

Beyond Nash, CE and ESS, game theory has explored other equilibrium concepts, including subgame perfect equilibrium, trembling hand perfect equilibrium, and proper equilibrium, among others. These refinements address limitations of the NE, particularly in dynamic and extensive-form games. We will only focus on CE and ESS in the current thesis.

Coordination and cooperation problems are fundamental challenges in social philosophy since Hobbes ([1650] 2016), and game theory has been an indispensable tool for tackling these problems due to its clarity and rigor.

* *Coordination problems* arise when individuals or groups need to choose between multiple possible equilibria, creating ambiguity about which solution will be selected. These problems are central to strategic interaction because they reflect situations where all parties would benefit from making compatible choices but may struggle to agree on a single option.
* *Cooperation problems*, on the other hand, highlight the conflict between individual rationality and collective benefit, where mutual cooperation yields a better outcome for all, but self-interest may lead to suboptimal results. Such challenges often require mechanisms to facilitate coordination or encourage cooperation, including social conventions or equilibrium selection techniques. Consequently, equilibrium concepts are fundamentally linked to coordination and cooperation problems because they model how rational agents arrive at stable solutions given others’ strategies.

Examples of coordination and cooperation problems include classic games like the Battle of the Sexes and the Prisoner’s Dilemma. In the former, a husband and a wife coordinate on choosing a leisure activity where everyone is satisfied with the choice, and in the latter, two prisoners independently either defect or cooperate with each other by uncovering their partner in crime to an officer. The payoff matrices of these games are shown below[[4]](#footnote-34).

$$
\begin{array}{|c|c|c|}
\hline
& Football & Ballet\\
\hline
Football & 2,1 & 0,0 \\
\hline
Ballet & 0,0 & 1,2 \\
\hline
\end{array}
\quad
\begin{array}{|c|c|c|}
\hline
& Cooperate & Defect \\
\hline
Cooperate & -1,-1 & -3,0 \\
\hline
Defect & 0,-3 & -2,-2 \\
\hline
\end{array}
$$

“Battle of the sexes” (left) and “Prisoner’s dilemma” payoff matrices

These matrices model real-world problems such as social dilemmas and negotiations. For instance, the Battle of the Sexes often represents situations where partners must choose between competing preferences, while the Prisoner’s Dilemma models the challenge of mutual cooperation versus self-interest in scenarios like arms races or public goods provision.

To illustrate the practical difference of equilibrium concepts in solving coordination problems, let us consider the Battle of the Sexes with pure Nash, mixed Nash and CE.

In pure Nash, two pure strategy equilibria exist: both players attend either Ballet or Football. These equilibria ensure perfect coordination but are inherently unfair, as one player always prefers the chosen event over the other.

A mixed strategy Nash equilibrium also exists, where players randomize their choices independently, but it risks miscoordination. Let the Husband choose Ballet with probability and Football with , and let the Wife choose Ballet with probability and Football with . Using the *indifference principle* according to which a player randomizes her strategies in a way that the opponent is indifferent between their own available strategies, we calculate probabilities:

1. For the Husband to be indifferent, the Wife’s mixed strategy must make his expected payoff from Ballet equal to that from Football:
2. For the Wife to be indifferent, the Husband’s mixed strategy must make her expected payoff from Ballet equal to that from Football:

Thus, in the mixed strategy Nash equilibrium:

* The **Husband** chooses Ballet with probability and Football with .
* The **Wife** chooses Ballet with probability and Football with .

The expected payoffs for both players in this equilibrium are:

* **Husband**:
* **Wife**:

This mixed strategy equilibrium represents a compromise balancing fairness and coordination through randomization, albeit less efficient than pure Nash equilibria due to inherent miscoordination risks[[5]](#footnote-35).

In contrast, CE utilizes public signals to coordinate actions effectively. For instance, a public signal such as a coin flip can recommend both players attend Ballet or Football equiprobably. This mechanism eliminates miscoordination and ensures equal expected payoffs for both players ( each). CE helps agents achieve higher payoffs and fairness compared to both pure and mixed Nash equilibria by leveraging shared randomness or communication.

To demonstrate how a signal affects the payoff structure, we add a new strategy *Follow Signal (FS)*, where players choose based on a fair coin flip (Heads = Ballet, Tails = Football). We can do this because CE is essentially a Nash equilibrium of a game augmented with an additional set of strategies (Gintis 2009b, 2009a). The payoffs here depend on actual coordination, not just expectations: we can calculate expected payoffs when one player uses and the other does not.

* **FS (H) vs Ballet (W)**:
  + Signal = Tails (50%): H chooses Football, W stays at Ballet →
  + Expected payoff:
* **FS (H) vs. Football (W)**:
  + Signal = Heads (50%): H chooses Ballet, W stays at Football →
  + Signal = Tails (50%): Both choose Football →
  + Expected payoff: .

Thus, the augmented game matrix becomes:

$$
\begin{array}{|l|c|c|c|}
\hline
& Ballet (W) & Football (W) & FS (W) \\
\hline
Ballet (H) & (2, 1) & (0, 0) & (1, 0.5) \\
\hline
Football (H) & (0, 0) & (1, 2) & (0.5, 1) \\
\hline
FS (H) & (1, 0.5) & (0.5, 1) & (1.5, 1.5) \\
\hline
\end{array}
$$

“Battle of the sexes” with correlated equilibrium

The strategy profile of represents a Nash equilibrium because neither player has an incentive to deviate. If a Husband switches to Ballet, he would only receive , a decrease from his current payoff of when the Wife remains at . Similarly, if the Wife switches to Football, she would receive only , a decrease from her current payoff of when the Man stays at . Since no profitable deviation exists for either player, the strategy profile is stable. Thus, the CE strategy is as an NE strategy of an augmented game. The difference is that CE are simpler to compute than NE and model real-world scenarios where external signals (e.g., traffic lights) guide decisions. In summary, CE expand the solution space of a game, offering improvements over Nash equilibria when players can leverage a coordination device.

Getting back to coordination problems, O’Connor (2019) distinguishes two classes of them:

* correlative problems (same choice to coordinate)
* complementary problems (different choices to coordinate)

In correlative coordination problems, agents need to converge on the same choice to coordinate successfully. For example, consider a driving game, where two players drive towards each other and each can choose the left or right side to drive on. If they both are on the same side and no one swerves, they might crash, and if each of them chooses a different side, they will stay safe. One important feature of this and other coordination problems is arbitrariness, meaning that it does not matter on what side both players would converge. Instead, what matters is that they either coordinate by choosing the same action, for example, swerving to the right.

On the game matrix, it is represented as two non-unique equilibria. It means that either of them solves the coordination problem.

$$
\begin{array}{|c|c|c|}
\hline
& Left & Right \\
\hline
Left & 1,1 & -1,-1 \\
\hline
Right & -1,-1 & 1,1 \\
\hline
\end{array}
$$

Driving game (correlative coordination game)

Complementary coordination problems, as opposed to correlative ones, require from agents different actions, or strategies, to coordinate successfully. As O’Connor (2019) points out, division of labor or resources is an example of this class of games. For instance, two roommates want to organize a party and invite guests. To proceed, they need to tidy up the house and order pizza delivery. If they both do the cleaning, there will be no food when the guests come, and if they both order pizza delivery, they will have plenty of food but be embarrassed by the mess at the house.

$$
\begin{array}{|c|c|c|}
\hline
& Order & Tidy \\
\hline
Order & -1,-1 & 1,1 \\
\hline
Tidy & 1,1 & -1,-1 \\
\hline
\end{array}
$$

Pizza game (complementary coordination)

The only difference between the two classes of coordination problems is either choosing same or different actions to coordinate successfully.

Coordination problems and conventions are intrinsically linked as former ones emerge when individuals or groups require aligned action for mutual benefit, necessitating communication and shared understanding to stabilize interactions. *Conventions function as a mechanism for predictable coordination by encapsulating mutual expectations*, thereby reducing ambiguity and establishing stable behavioral patterns within a social context. David Lewis’s theory of conventions as coordination equilibria, explored in the subsequent section, provides a central treatment of this relationship.

## Intellectual influences of Lewis’s “Convention”

The intellectual atmosphere in which Lewis’s *Convention* was developed was mostly engaged with questions of language, meaning, and social behavior. Several intellectual movements and concerns shaped the development of his theory.

In the mid-20th century, the interest in influence of social practice on linguistic meaning kept growing, as philosophers like Quine (**quine1960?**) and Wittgenstein (**wittgenstein?**) argued that meaning arises from shared use within a community. Wittgenstein highlighted that language’s meaning emerges through public usage, rather than inherent semantic properties. For instance, “game” has no fixed definition but derives its meaning from the activities associated with it. Building on this tradition, Lewis sought to explain how linguistic conventions form, stabilize, and persist in communities by providing a systematic account of their development over time. By conceptualizing meaning as coordinated behavior, Lewis laid a foundation for viewing language as a socially orchestrated activity rather than an innate or purely individualistic construct. Consequently, communication relies not on objective meanings but on mutual expectations about usage, emphasizing convention’s crucial role in language (Lewis [1969] 2008).

The Zeitgeist of analytic philosophy in the 1960s grappled with the legacy of Logical Positivism, which, through formal logic and empirical verification, defined meaning based on analytically true statements or verifiable empirical claims (Godfrey-Smith 2003). However, by the 1960s, critiques from Quine, Putnam, and others challenged this framework, particularly the distinction of analytic/synthetic truths, the former being true in virtue of their meaning and the latter in virtue of their relationship to the world.

Quine rejected traditional notions of necessity and analyticity, asserting ontological commitments are embedded within theories and language (**quine1951?**; **quine1960?**; **quine1969?**), emphasizing empirical evidence and pragmatic considerations in shaping beliefs. His critique of analyticity underscored the revisability of language, highlighting conventions as mutable rather than fixed. Putnam’s “Twin Earth” thought experiment[[6]](#footnote-37) further developed these ideas, advocating semantic externalism—the view that word meaning depends on external facts, not solely on mental states—challenging internalist accounts of meaning and emphasizing the role of external factors in linguistic practices. Consequently, conventions are understood as influenced by contextual and environmental factors, moving beyond purely internal or necessary determinations.

Lewis’s theory of convention was a way to address this intellectual shift by emphasizing the contingent nature of meaning. Rather than being dictated by any necessity, conventions arise as arbitrary but stable solutions to coordination problems, reflecting a more pragmatic and flexible understanding of linguistic meaning and social practices. It highlights that even the most strict customs started as contingent behavioral patterns which might have been otherwise but have been amplified with each iteration. This perspective is deeply rooted in Quinean ideas about language as being subject to revision, adaptation, and negotiation within a community or culture.

Another major philosophical concern that Lewis addressed was the ontology of social rules and norms, profoundly influenced by Hume’s work. Lewis developed Hume’s idea of conventions emerging and persisting even in the absence of centralized enforcement. Lewis argued that conventions are self-reinforcing: once established, individuals have no reason to deviate as long as others continue to conform. The major deviation from Hume’s thought was an accent on rationality of agents as the source of such conformity, whereas Hume emphasized psychological custom.

An example of this can be seen in the development of money as a medium of exchange. Initially, various objects like cattle, shells or metal coins served as currency. Over time, paper money became widely accepted, not because of any intrinsic value, but because people expected others to accept it in transactions. This insight was later influential in discussions of spontaneous order and decentralized systems in political philosophy and economics, particularly in the work of Hayek (**hayek1973?**). By explaining conventions as natural outcomes of repeated social interactions, Lewis contributed to a broader understanding of how norms, institutions, and linguistic practices can arise organically without explicit design or coercion.

Furthermore, Hume’s skepticism about moral realism, a position stating that objective moral norms exist, played a role in shaping Lewis’s view of conventions as arbitrary yet stable[[7]](#footnote-38). Hume argued that moral distinctions are not grounded in objective properties but in human sentiment and social conditioning. Similarly, Lewis contened that conventions are not determined by any intrinsic necessity but arise contingently through social practices. For instance, the choice of driving on the right or left side of the road is arbitrary, yet once established, it becomes self-reinforcing because all individuals benefit from adherence to the convention. This reflects Hume’s broader thesis that social order emerges not from absolute principles but from shared expectations and learned behaviors.

If the problems of meaning, language and conventionality served as the issue Lewis wanted to attack and Hume’s notion of convention was resource to build upon, he still needed a tool to construct his argument with. He found it in game theory (**vonneumann1944?**) and, in particular, in Schelling’s approach to strategic interaction in “mixed motive” games (Schelling 1980).

Game theory offered a structured mathematical framework for analyzing strategic interactions among individuals conceived as rational actors. Lewis’s engagement with game theory and decision theory was facilitated by this prevailing intellectual trend. The emphasis on formal models and rational choice provided a common language and conceptual framework for discussing social behavior across diverse disciplines, making it a natural progression for a philosopher like Lewis to explore these powerful analytical tools in his own work.

Schelling’s work represented a significant departure from prevailing game theory’s emphasis on zero-sum conflict (when there is always a winner and a loser), recognizing that real-world interactions frequently exhibit “mixed motives” or simultaneous conflicting and converging interests. He critiqued the limitations of purely mathematical analysis of strategic interaction and advocated for empirical research to illuminate the conditions shaping behavior, specifically considering opportunities for communication and the presence of attractive alternatives. This expanded scope featuring both conflict and cooperation included the very phenomena of cooperation and coordination that drew Lewis’s attention in the context of the problem of social conventions.

Schelling argued that conflict and cooperation are not necessarily opposing forces but are deeply intertwined in strategic interactions. One of his key contributions was the concept of *credible commitment*, where the ability to commit to a particular strategy in advance can influence an opponent’s decisions (**schelling1960?**). A fundamental aspect of this is *self-binding*, where a player deliberately restricts her own options to strengthen bargaining position.

Another crucial insight was the concept of *focal points* (also known as Schelling points), which are solutions that individuals naturally gravitate toward in coordination games without explicit communication. Schelling demonstrated this through experiments where participants, when asked to choose a meeting place in New York City without coordination, overwhelmingly selected noon at Grand Central Terminal, although it was a location with no inherent payoff advantage but high cultural prominence (**schelling1960?**).

In the study of *pure coordination games*, Schelling examined interactions where players share interests but lack communication, such as selecting matching integers for a reward. Participants often converged on salient choices, such as the number 1, due to its distinctiveness as the smallest positive integer (**schelling1960?**). His work also refined the Nash equilibrium by demonstrating how focal points can help identify stable and salient outcomes among multiple NE (Lewis [1969] 2008, 78). Furthermore, for conflict scenarios, he introduced the concept of *“threats that leave something to chance”*, showing that probabilistic threats, such as partial mobilization, can deter adversaries more effectively than deterministic ones by leveraging uncertainty to maintain deterrence (**schelling1960?**).

Lewis formalized Schelling’s insights into a theory of conventions, defining them as solutions to recurrent coordination problems where agents align on focal points due to mutual expectations (Lewis [1969] 2008, 43). Conventions rely on extrinsic incentives, such as avoiding coordination failure, rather than intrinsic obligations. Lewis also emphasized that communication itself is a coordination game, where signals derive meaning from shared conventions (Lewis [1969] 2008, 95).

One of the central ideas Lewis took from Schelling is the concept of focal point, or salience. He showed that social conventions arise as focal points for coordination. For instance, in many societies, people drive on one designated side of the road not because of an inherent preference for that side, but because universal adherence to a single convention ensures safety and predictability. Building on that idea, Lewis argues that agents select the most salient convention which “stands out” from alternatives, either through precedent, explicit agreement, or intrinsic properties. According to Lewis, salience, a subjective psychological trait independent of the strategic situation, governs convention emergence and conformity. Specifically, Lewis addresses how conventions arise (dynamics – through initial selection and subsequent salience amplification) and why people conform (statics – due to the overwhelming salience of a pre-existing convention, fostering an expectation of adherence). Subsequent refinements of Lewis’s theory reimagine and formalize the notion of salience mostly through evolutionary lens Skyrms (2014a).

Another crucial concept Lewis adopts from Schelling is the role of expectation and self-enforcement in strategic equilibrium. Schelling showed that in many coordination scenarios, once an equilibrium is established, deviation becomes irrational since the costs of uncoordinated action outweigh potential individual gains. Lewis builds on this by defining conventions as self-perpetuating: once a convention is in place, individuals follow it not because of external enforcement, but because mutual expectations make deviation costly. This is evident in linguistic conventions, where the use of certain words and grammatical structures persists because everyone expects others to conform to them.

Furthermore, Lewis’s notion of *common knowledge*, foundational to his theory of conventions, derives from Schelling’s emphasis on mutual awareness within strategic interaction which is tightly connected with salience. Though Schelling lacked formalization, he highlighted the crucial role of shared understanding for successful coordination. Lewis expanded upon this, asserting that convention stability necessitates not just adherence, but also recognition as the expected behavior within a group, thereby enabling convention maintenance across large populations.

By drawing on Schelling’s work, Lewis was able to provide a game-theoretic foundation for the study of conventions, demonstrating how they emerge, stabilize, and persist over time. Whereas Schelling’s focus was on strategic choices in conflict and negotiation, Lewis extended these principles to the domain of language, social norms, and epistemic coordination, thus broadening the applicability of game-theoretic insights to philosophy and social science. As a result, Schelling’s *The Strategy of Conflict* remains one of the key intellectual influences behind Lewis’s *Convention* and its enduring impact on theories of social coordination.

## Lewis’s theory of conventions

Lewis defined social convention as an arbitrary yet self-sustaining behavioral pattern emerging from repeated coordination problems between two or more players. Its distinctive feature is players’ conformity to these behavioral patterns, for they expect others to do so, and it is *common knowledge* that every player is expected to conform. Deviation from a conventional choice of action leads to lower payoff, so players do not have incentives to deviate unilaterally which is on their own. For example, if everyone drives on the right side of the road, it is rational for each driver to do the same to avoid collisions. Lewis (Lewis [1969] 2008, 76) formulates convention as follows:

A behavioral regularity within a population in a repeated situation qualifies as a convention if and only if:

1. Every member of conforms to .
2. Each individual expects others to conform to .
3. All members have similar preferences regarding possible behavioral patterns.
4. Each person prefers universal conformity to , provided that nearly everyone else adheres to it.
5. Members would also prefer an alternative regularity under the same conditions, as long as and are mutually exclusive.

Lewis later refined his analysis to accommodate occasional deviations from convention and Skyrms (2023) recently introduced *quasi-conventions* as unstable conventions based on yet another concept of *coarse correlated equilibrium*. Despite this, much of the academic discourse focuses on the strict version of Lewis’s definition.

Lewisian convention is a special kind of equilibrium called *coordination equilibrium*, which roughly resembles NE, but extends beyond it. In NE, no participant can improve their outcome by unilaterally changing their strategy. If deviation strictly reduces payoff, the equilibrium is considered strict. In this sense, NE represents a “steady state,” where each individual acts optimally given the actions of others. However, Lewisian convention extends beyond NE by emphasizing collective preference for conformity, even when minor deviations occur.

Lewis’s framework highlights *arbitrariness* in conventions, where is defined as a convention only if an alternative could serve equally well. This acknowledges that conventions are contingent choices among possible solutions rather than inherent necessities which continues the insights of Quine (**quine1969?**), Putnam (**putnam1975?**) and others.

Additionally, Lewis introduced the concept of *common knowledge* and made it a condition for a regularity to be a convention, where a fact is common knowledge if:

* Everyone knows ;
* Everyone knows that everyone knows ,
* Everyone knows that everyone knows that everyone knows , and so on.

This recursive understanding of knowledge has spurred extensive discussion in both philosophical and game-theoretic literature. Aumann (Aumann 1976) and Schiffer (Schiffer 1972) have developed formalizations of common knowledge, diverging from Lewis’s original informal approach.

As we will tour this and other aspects of Lewis’s theory in detail later in this chapter, it suffices to mention that further reception of his theory saw the common knowledge requirement too cognitively demanding and unrealistic (Gilbert 1992; Binmore 2008; **camerer2003?**; Bicchieri 2005; P. Vanderschraaf 1998).

As Lewis’s theory uses game theory, rationality plays a fundamental role in his framework. Lewis assumed that agents are instrumentally rational, meaning they choose actions that maximize their expected utility given their beliefs and expectations about the world and the behavior of others. Although the entire metaphor of humans as maximizing agents has been questioned (Cédric Paternotte 2020), it still serves as guidance in economic theory (Gintis 2007a; Gintis and Helbing 2013), biology (S. Okasha 2017; Samir Okasha and Binmore 2012; Engel and Singer 2008) and human ecology (Mouden et al. 2012; Sterelny 2012a). However, there are alternative views on the requirement of agent’s rationality for conventions to exist. Ruth G. Millikan (2022) suggests that conventions stabilize only by the weight of precedent, thus not requiring any rationality or consciousness.

Lewis’s notion of conventions weaves behavior, beliefs, preferences, and expectations into a framework of common knowledge and rationality to explain the stability of conventions. Each part of the definition is vital:

* common knowledge ensures a shared understanding of the convention,
* the preference for conformity incentivizes adherence given others’ cooperation,
* rationality guides individual choices within the context of shared expectations.

As a primary motivation for Lewis’s analysis was to address the philosophical problem of linguistic meaning, he aimed to argue that language is grounded in conventions which do not require up-front agreement on terms. Just as drivers coordinate on a side to drive without a formal contract, speakers of a language develop conventions of using sounds or gestures to refer to specific things through repeated interaction and mutual expectations. Lewis viewed language as a system of signaling, where meaning arises from the conventional association between signals (words, phrases) and states of the world. For example, the word “cat” conventionally signals the presence of a feline. This convention is sustained because speakers generally intend to be truthful and listeners generally trust that they are being told the truth. This mutual expectation and reliance on the regularity of signal-meaning pairings allows for effective communication, which is a form of coordination.

This led Lewis to delineate *behavioral* and *signaling* conventions (**lewis2008?**), where the former coordinate actions and the former coordinate meaning in communication. As a prototypical example of a signaling convention, Lewis gives a story of Paul Revere and the lanterns hung in the steeple of the Old North Church used to warn colonial militia about approaching British Troops in 1775. Two hung lanterns conveyed that troops are advancing by sea, one by land. Additionally, the actions of a message’s receiver, given each of these signals, would differ. Senders and receivers of a message coordinate on following a pre-established pattern of “if X, do Y” like in the example with lanterns[[8]](#footnote-40).

For Lewis, signaling conventions are a special case, or a subclass, of behavioral conventions as they share basic properties like arbitrariness, conformity and being common knowledge. Signaling conventions differ in that they involve communication and interpretation of meaning and solve coordination problems by *information transfer*. They require encoding/decoding which is producing and interpreting signals.

An important feature of the relationship between these two classes of conventions is that, according to Lewis, signaling conventions fundamentally rely upon and are shaped by pre-existing behavioral conventions. For example, language meanings of words depend on both parties’ adherence to established norms of pronunciation and grammar. Signaling systems frequently exhibit nesting, where specific conventions are embedded within larger behavioral regularities. For instance, raising one’s hand to speak during a meeting is a signaling conventions nested within a broader behavioral convention of turn-taking (Peter Vanderschraaf and Skyrms, n.d.).

There is a formal distinction between behavioral, or “general” as Lewis call it, and signaling conventions. In signaling games, the players can be either senders or receivers, where the former owns private information about the world state and send a signal about it and the latter observes the signal and acts on it. More formally, it looks like the following:

1. **World states**: L (left) and R (right)
2. **Signals**: V₁ and V₂
3. **Actions**: Aᴸ (left action) and Aᴿ (right action)

| Role | Strategy | Description |
| --- | --- | --- |
| **Sender** | S₁ | Signal V₁ if L, V₂ if R |
|  | S₂ | Signal V₂ if L, V₁ if R |
| **Receiver** | R₁ | Choose Aᴸ if V₁, Aᴿ if V₂ |
|  | R₂ | Choose Aᴿ if V₁, Aᴸ if V₂ |

And in a matrix form it looks as follows:

$$
\begin{array}{|c|c|c|}
\hline
& R₁ & R₂ \\
\hline
S₁ & (1,1) & (0,0) \\
\hline
S₂ & (0,0) & (1,1) \\
\hline
\end{array}
$$

Signaling game

If a sender’s signal representing a world state is correctly acted upon by receiver, both parties get the payoff of and if either party fails to map (“encode” or “decode”) information, they get . There is a plethora of possible options within this informational “layer” of signaling system like pooling, sysnonyms, deception and others extensively studied primarily by philosophers of biology (Skyrms 2010b, 2010a; Huttegger and Skyrms 2008; Godfrey-Smith 1991; Shea, Godfrey-Smith, and Cao 2018; Martínez 2019).

Godfrey-Smith (2014) refined Lewis’s model by distinguishing *state-act* and *act-act* coordination:

* *State-Act* coordination: signals map states to receiver action;
* *Act-Act* coordination: signals synchronize action between agents without any external events.

*Act-act* coordination allows to view Hume’s boat rowers as an act-act signaling system: the rowers’ rhythmic strokes serve as imperative signals (“Row now!”) that directly coordinate mutual actions rather than conveying information about external conditions (Martínez and Godfrey-Smith 2016). The absence of an exogenous state reduces the system to a pure coordination game employing Nash or coordination equilibrium, where the “signal” (stroke rhythm) functions as a self-reinforcing convention stabilized by common interest and reciprocal expectations. Unlike state-dependent signaling of *state-act* coordination, which requires alignment between acts and external facts, *act-act* systems like the rowboat prioritize *interpersonal synchronization* through real-time behavioral feedback, illustrating how communication can organize joint action without representational content.

A paradigmatic real-world example of a *state-act* signaling system is alarm calls specific for each type of predator. For example, vervet monkeys have a call for seeing eagles which conveys hiding in the grass and a call for seeing a snake conveying climbing on a tree (Seyfarth and Cheney 1990). A perfect connection between a world state, signal and action comprises a signaling system.

Although formally similar, as both behavioral and signaling conventions can be described as games with players and payoffs, they differ in that the latter have an additional “layer” of information between players. And although Lewis himself proclaimed signaling conventions a subcategory of behavioral ones, the relationship between them is not clear. For Skyrms, signals *inform* action, and signaling networks *coordinate* action, which implicitly conveys signaling conventions as underpinning behavioral ones and nit vice versa. Skyrms further suggests that signaling is responsible for the evolution of teamwork itself (Skyrms 2010a), which questions Lewis’s hierarchical categorization and creates a version of a chicken-and-egg problem, which is out of scope of this thesis.

## Criticisms and generated problems of Lewis’s theory

Lewis’s theory has been criticized on many grounds, and, as Rescorla (2024) notes, virtually every component of his theory has been under attack: from imprecise notion of equilibrium concept to the very necessity of conventions for solving coordination problems. However, many initial criticisms have been met in the refinements and extensions of Lewis’s theory by later scholars.

There are five main areas of criticism of Lewis’s account of conventions:

1. conformity requirement and hidden normativity
2. overestimation of arbitrariness
3. common knowledge requirement and source of salience
4. connection between conventions and coordination problems
5. imprecise equilibrium concept

We will survey 1-4 here as 5 is an extension rather than critique which we will address in the next section on refinements of Lewis’s theory. Each subsection starts with immediate criticism of Lewis’s theory and continues with a larger problem Lewis’s theory generated or contributed to.

### Hidden normativity of conventions

One of the major criticisms of Lewis’s theory of conventions is unrealistic conformity requirement expressed of his 4-th clause: “each person prefers universal conformity to , provided that nearly everyone else adheres to it”. As some scholars points out, this strict requirement rules out such regularities as sending thank-you notes after dinner as non-conventional, for they do not require complete conformity (Gilbert 1992). Many commentators find this unintuitive as we usually call any mutually expected behavioral regularity a convention regardless of its level of conformity.

However, a possible defense of Lewis’s position is to restrict a social group where convention takes place and to add that “each person *within a bounded social group* prefers universal conformity to …”. This addition addresses Gilbert’s criticism in that it supports an idea of near-complete conformity relative to the scale and size of a social group with operative convention. If sending thank-you notes after a dinner within a certain group is indeed a convention, not writing such a note would at least disappoint a dinner host. Of course, this might not impose any external sanctions on a guest not writing a thank-you note. However, conformity relative to group size highlights inherent normativity in the form close to normative expectation, which Bicchieri (2005) considers an essential ingredient of social norms rather than conventions.

As can be seen, a convention’s level of conformity poses a deeper problem—normativity of conventions. The level of conformity helps distinguish between conventions as regularities *de facto* and *de jure* (Rescorla 2024), where the former describe actual behavior and the latter prescribe how individuals should behave in certain situations. Lewis himself anticipated such objections and claimed that conventions eventually become social norms. This claim later generated a major controversy over the relationship between those (Bicchieri, Muldoon, and Sontuoso 2023). Level of conformity points to a problem of the source of such conformity, which concerns Lewis’s critics.

Gilbert (1992) contends that Lewis’s account ignores the normative force of conventions. For Gilbert, the source of conformity of conventions is *joint commitments* that bind participants to collective ends, creating obligations to conform and justifying criticism of defectors. Lewis’s model, which reduces conventions to equilibrium strategies in coordination games, cannot explain why individuals feel *obliged* to comply with conventions (e.g., stopping at red lights) or apologize for breaching them. Gilbert also disputes Lewis’s focus on coordination problems, arguing that many conventions like etiquette rules lack clear coordination benefits and instead reflect shared commitments. In Lewis, there is appeal to instrumental rationality which maximizes expected value and avoids sanctions, but some scholars see this as insufficient to substantiate conventions.

For instance, Guala and Mittone (2010) study the extent to which Lewis conventions are normative. They address both theoretical and empirical aspects and conclude that Lewis has put forward a scientific theory of conventions and not an analysis of the folk notion, and that conventions do indeed have intrinsic normativity beyond that of instrumental rationality. However, there is another strand of scholars that disagree and put forward that the only normativity of conventions is that of instrumental rationality (Gold and Sugden 2007; Bacharach and Bernasconi 1997).

As Guala and Mittone (2010) report, in experimental settings, given an iterated Ultimatum or Prisoner dilemma game, only 29% of potential deviants in the lab choose to breach emergent convention. Players in these experiments may unintentionally create extra pressure to conform with their shared history of action, beyond the requirements of rational decision-making and social norms. However, the exact mechanism of additional normative expectation formation is to be discovered.

From this, Guala concludes that Lewis’ model provides an incomplete account of conventions’ ontology. Data suggest that Lewisian conventions acquire normative force through repeated play, and any future model must account for this. This has non-trivial implications for theory and practice, as it implies that habits and customs may be hard to disrupt.

In a similar vein, F. Hindriks (2019) claims that instrumental rationality cannot motivate adherence to conventions and norms and their perception as legitimate. Instrumental rationality with its costs and expected utilities fails to capture the motivation by the normative part of convention itself and not by the costs of its violation. Hindriks claims that it is *normative expectations* and *normative beliefs* that complement sanctions as a source for norm existence and perception as legitimate.

In a broader context, the problem of normativity of Lewis conventions ignited a lively debate in philosophy regarding the relationship between conventions and social norms. They share a foundation in regularities and mutual expectations, but diverge in normative force, enforcement mechanisms, and social functions. Lewis’s model offers a minimalist, rationalist account of coordination, whereas social norms represent a richer, more complex landscape of social regulation, deeply shaped by values, sanctions, and cultural meanings. Thus, norms are seen as conventions acquiring enforcement capacity beyond pure coordination with repeated play.

Ullmann-Margalit (1977) gave one of the first game-theoretic conceptualizations of social norms. She argued that, as groups interact repeatedly, they develop expectations based on both obligation and the threat of social sanctions. These expectations go beyond simple conventions, relying on individuals’ internalized sense of duty and the potential for consequences like disapproval or ostracism. According to Ullmann-Margalit, this represents a move from coordinating behavior solely through rational self-interest to one where compliance is motivated by a deeply held conviction.

More precisely, Ullmann-Margalit emphasized normative authority as a mechanism of norm emergence, where they arise as informal solutions to coordination, cooperation, and mixed-motive problems characterized by divergent individual and collective interests. She conceptualized a *history space* as sequences of past interactions, where each history pairs joint action profiles with observed sanction signals (negative, neutral, positive). A *normative choice function*

maps each history to a probability distribution over actions, reflecting both *strategic expectations* and *normative weights*.

Then, for any history , an agent’s expected utility from choosing action is given by:

where:

* is the material payoff based on beliefs about opponents’ strategies inferred from .
* is the *internalized norm intensity*, capturing both social and psychological sanctions.
* scales the importance of normative pressures relative to material gains.

Ullmann‑Margalit further suggests that norm intensity evolves according to a *reinforcement learning* update:

where is the average sanction observed when action was taken at time , and is a sensitivity parameter. Over time, this embeds the *frequency and severity* of social approval or disapproval into agents’ preference structures.

A critical feature of Ullmann‑Margalit’s sketch is the *threshold effect*: once exceeds a context-dependent threshold , normative considerations dominate, and agents will comply even against short-term material incentives. Formally, if

then places all weight on , the *normatively prescribed* action, regardless of smaller material gains from deviation.

This model suggests path-dependence as early interaction profiles and sanction signals can lock a community into a particular normative equilibrium, making norm shifts resistant to modest perturbations. In addition, in public goods or trust games, where favors free‑riding, a sufficiently high can sustain cooperation via internalized guilt and peer sanctions. Crawford and Ostrom (1995) later modeled a similar situation of cooperation problems with -parameters seen as additionally incurred costs in the form of potential sanctions, which, being sufficiently high, can transform a cooperation game into a coordination game. 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. This shows that normative rules can be coordination devices, or “choreographers”, as Gintis puts it (Gintis 2009b).

Ullmann‑Margalit also acknowledged multiple channels like formal penalties, gossip and emotional costs that feed into , allowing norms to persist even when formal institutions are weak. Through these mechanisms, Ullmann‑Margalit transformed Lewis’s static, expectation-based model into a dynamic framework that accounts for the emergence, stability, and transformations of social norms with genuine normative force.

Although not directly related to Lewis’s conventions, the pioneering work of J. M. Epstein and Axtell (1996) in generative social science employed agent-based modeling to investigate the spontaneous emergence of social conventions and norms, driven by endogenous norm formation without centralized enforcement. It is similar to Lewis’s project and is referenced here as a proof-of-concept that norms can be “grown” from behavioral pattern a la Lewis.

Rooted in agent-based computational economics and evolutionary game theory, Epstein’s models utilize bounded rationality, with agents operating under simple rules and limited information. This rejects traditional assumptions of perfect rationality or equilibrium. System-level behavior is analytically intractable, necessitating simulation as a primary tool. This framework aligns with game theory, specifically repeated and evolutionary games, where payoffs are contingent on frequent population-level strategies, and norms emerge as persistent patterns of behavior sustained by costly deviation or social punishment.

Epstein’s models, utilizing ‘Sugarscape’ simulations, employ heterogeneous agents governed by dynamic, simple behavioral rules, independently generating both global conventions and localized norms without centralized coordination. These models highlight the critical roles of path dependence – the influence of past decisions on future behavior – and network topology in shaping social pattern stability. Translating Lewis’s coordination concept into a computational framework, Epstein demonstrated how agent diversity, stochastic processes, and adaptive learning mechanisms can produce durable social regularities exhibiting normative implications.

In Epstein’s canonical model, agents occupy a grid and interact locally. Each agent follows a simple rule-based strategy and can observe the behavior of its neighbors. Over time, they update their behavior based on local observations and possibly some payoff structure.

The theoretical core of Epstein’s model can be expressed using a **game-theoretic payoff matrix**, where agents derive utility from conforming to or deviating from a local behavioral pattern:

Let the strategy set be , representing **conforming** and **deviating** . The payoff matrix for an agent may be:

Where:

* : reward for mutual conformity
* : sucker’s payoff (you conform, others don’t)
* : temptation to deviate
* : punishment for mutual deviation

Agents update their strategy based on *success-based imitation* (copying the strategy of the most successful neighbor), or using rules akin to *replicator dynamics* in evolutionary game theory:

Where:

* : proportion of agents using strategy
* : payoff matrix
* : strategy distribution vector

However, in Epstein’s ABM, this is implemented *discretely and locally*: agents switch strategies probabilistically based on their relative payoffs and observed neighborhood norms.

Over time, consistent behavioral patterns like norms may emerge. These norms are not imposed but *self-organized* through repeated interactions and adaptation. A key insight from Epstein’s work is that *norms can be sustained through decentralized enforcement*, such as local peer pressure or reputation systems. For instance, agents might adopt a punishment rule: if a neighbor deviates from a prevailing norm, the agent may impose a cost on them (e.g., by refusing to cooperate in future interactions). This introduces *meta-norms* (norms about enforcing norms), which further stabilize the system.

Several important insights emerged from Epstein’s simulations:

* *Multiple equilibria*: Different regions of the simulation space can stabilize on different norms, illustrating the *path-dependence* of social systems.
* *Robustness*: Norms are robust to noise and perturbations, depending on enforcement mechanisms.
* *Phase transitions*: Under certain conditions, the system exhibits critical behavior where small changes in parameters (e.g., punishment cost) can lead to large-scale shifts in norm adoption.

Joshua Epstein’s work bridges the gap between micro-level behavioral rules and macro-level social patterns. By embedding game-theoretic interactions in a spatially explicit agent-based model, he demonstrated how complex social norms can emerge from simple local interactions. His simulations provide a powerful tool for exploring the dynamics of social norm formation, challenging the view that norms require centralized design or enforcement. Instead, they may be the natural outcome of decentralized, strategic adaptation in social environments.

Another treatment of normativity in regard to conventions is due to Young (1998). Young’s theory of social norms represents a major step in formalizing how collective behavioral patterns arise and stabilize in decentralized populations. His key contribution lies in drawing a conceptual and mathematical distinction between *conventions* and *social norms*, and in showing how norms can emerge and persist even when they are not individually payoff-maximizing. While conventions are defined as self-enforcing equilibria of coordination games, social norms in Young’s account involve a deeper layer of *prescriptive expectations* and *enforcement structures*, including social punishment for deviation (Young 1998, 83–85). This distinction allows him to move beyond the Lewisian model of convention (Lewis [1969] 2008) and more closely approximate real-world practices like tipping, queuing, or property respect, which often persist even in the absence of external enforcement or explicit payoff dominance.

Young modeled social behavior as an evolutionary process grounded in *adaptive learning*. Agents are situated in a population and repeatedly engage in *local interaction games*, where strategies are updated according to boundedly rational rules such as *myopic (or short-sighted) best response* focusing on the present payoff possibilities. Importantly, agents are assumed to make occasional errors, either due to experimentation or misperception which introduces *stochasticity*, or non-deterministic randomness, into the learning process. This leads to a Markov chain[^Markov=chain] over the space of population states, where each state corresponds to a configuration of strategies across individuals.

To analyze the long-run behavior of such systems, Young introduces the concept of *stochastic stability*. Let be the finite set of population states, and let be a transition probability matrix parameterized by a noise level . Then the *stationary distribution* satisfies:

and the set of *stochastically stable states* is defined as:

To compute , Young uses the *resistance tree* method. Each transition between states is assigned a *resistance* based on how improbable it is (i.e., how many agents must err simultaneously), and stochastically stable states are those with the *lowest total resistance* across spanning trees rooted at those states (Young 1998, 103–7).

This framework has several far-reaching implications. First, it provides a *selection principle* for equilibrium in games with multiple Nash equilibria. Unlike refinements based on rationality assumptions, Young’s selection is *behavioral and dynamic*, requiring only minimal cognition. Second, it allows the modeling of *social enforcement*: norms can be represented as equilibria of games with endogenous punishment strategies, wherein agents prefer to punish norm violations even at a personal cost (Young 1998, 90–92). These punishment mechanisms can themselves be stochastically stable, giving rise to *self-enforcing normative structures* without assuming rational foresight or institutional backing.

Young’s framework in also richly connected to other theoretical frameworks. Young’s approach aligns in spirit with Skyrms’s view of *local interaction as a source of correlation* in the emergence of conventions (**skyrms1996?**), but differs in emphasis. While Skyrms treats conventions as emerging from *evolutionary dynamics with minimal structure*, Young introduces *explicit sanctioning mechanisms and social expectations*, making norms qualitatively distinct from conventions. Moreover, Young’s stochastic stability framework offers a *non-epistemic interpretation* of equilibrium selection that contrasts with the *Bayesian rationality* underlying correlated equilibrium in Aumann’s model (Aumann 1974). In Aumann’s framework, a *correlation device* (e.g., a public signal) coordinates beliefs to yield a correlated equilibrium, relying on agents’ rational Bayesian updating. Young’s framework, by contrast, explains the emergence of coordination *without any epistemic content*, driven instead by population-level statistical stability.

In sum, Young’s theory provides a mathematically rigorous account of how *prescriptive social norms*, backed by informal sanctions, emerge and persist in large populations through *local adaptation and noise-driven selection*. His work not only bridges the gap between evolutionary game theory and institutional analysis, but also supplies tools for identifying which norms are likely to prevail under given interaction structures and stochastic conditions.

Bicchieri (2005) offered a distinct and now mainstream epistemic and psychological approach to social norms departing from traditional game-theoretic, functionalist, and purely behavioral models. She distinguished conventions from social norms based on *conditional preference*. According to her, individuals follow a rule if they:

1. expect sufficiently many others to follow it (empirical expectation)
2. believe sufficiently many others think they ought to follow it (normative expectation).

In Lewis’s conventions, only empirical expectations matter while normative expectations and potential sanctions are absent. By contrast, actual norms depend on both layers of expectation and are upheld through social sanctions, external (punishments) or internal (guilt) (**bicchieri2006?**). Bicchieri’s framework highlights that modifying agents’ beliefs about others’ attitudes can transform a convention into a norm, offering actionable insights for policy and institutional design.

Bicchieri posits that a behavioral rule is a social norm in a population with respect to a situation if:

1. The rule is *recognized* to apply in situation .
2. A sufficiently large subset of conforms to in , and each individual has the *empirical expectation* that others conform to in .
3. Each believes that others think she ought to conform to in (i.e., has *normative expectations*).
4. Each has a preference to conform to *on condition* that:
   1. she believes others conform to in ;
   2. she believes others believe she ought to conform to in .

This is often expressed via a *preference conditional on expectations*:

where:

* = empirical expectation: “Others will do in ”
* = normative expectation: “Others expect me to do in ”

This conditionality sets Bicchieri apart from rational choice or equilibrium-based models. For her, compliance with norms depends not on fixed utility functions but on *context-sensitive preference shifts* driven by belief structures.

Contra Lewis, Bicchieri insists that *norms are not reducible to equilibrium strategies* or to mutually expected behavior. They involve a prescriptive layer comprised of what others think one *ought* to do. Moreover, unlike Lewis, she does not require common knowledge of the regularity. Instead, norms can exist in mixed populations and be locally activated (**bicchieri2006?**).

Bicchieri argues that norm compliance is largely situational and dependent on expectation structures, not internalized values or long-term moral training like *self-command* in other frameworks adjacent to game theory (Elster 1989). Her experimental evidence suggests that many individuals are *conditional cooperators*, responsive to perceived expectations, and prone to *strategic norm evasion* when the normative context is weak (**bicchieri2006?**). Norms, for her, do not require full internalization but are highly contextual and can be followed instrumentally in appropriate contexts.

Compared to Ullmann-Margalit (1977), who conceptualized norms primarily as solutions to recurrent social problems like coordination, cooperation and avoidance of conflict with norms as functional entities, Bicchieri’s view departs in two ways:

1. Bicchieri decouples norm emergence from functional necessity meaning that not all norms solve problems, and some persist even when suboptimal
2. She builds a *bottom-up epistemic model*: a norm exists not because it solves a coordination problem, but because agents *believe* it exists, and condition their preferences accordingly. The existence of a norm is thus a *psychological fact about belief networks*, not merely a systemic solution.

Bicchieri thus has offered a model of norms as *contextually activated scripts*, grounded in micro-level epistemic structures, and capable of explaining variability, fragility, and rapid change in social behavior. It is also important to note that Bicchieri’s agents are entirely conditional, meaning they form, have and update explicit beliefs and have *cognitive schemata* sufficient for processing these beliefs (Bicchieri and McNally 2018).

One more influential conceptualization of social norms is due to Gintis who offered a *multi-level evolutionary account* of social norms that integrates insights from game theory, behavioral economics, evolutionary biology, and complex systems theory. Unlike approaches that treat norms either as equilibrium strategies (Lewis) or as epistemic constructs (Bicchieri), Gintis argued that *norms are a form of socially transmitted rule-based behavior* that co-evolves with the human capacity for cooperation and punishment, and whose persistence is explained through *gene–culture coevolution* (**gintis2003?**; Gintis 2009a).

Gintis defined a norm as a rule of behavior that is:

1. *Universally shared* within a reference group,
2. *Individually internalized*, so that deviation provokes negative emotions like guilt or shame,
3. *Enforced through third-party punishment*, and
4. *Costly to individuals*, yet *adaptive* at the group level (**gintis2003?**).

The evolutionary viability of such norms arises from the interplay between individual fitness and group selection: although norm-followers may incur costs, groups with strong norm adherence—especially norms of cooperation, fairness, or punishment—outperform less cohesive groups in intergroup competition. This is formalized in models of *multi-level selection*, where within-group dynamics favor selfishness, but *between-group dynamics favor cooperation* mediated by norms. As Vlerick (2019) suggests, solutions to coordination problems emerge from within-group dynamics, while solutions to competition ones are largely selected through between-group competition. Within-group dynamics explain why salient coordination rules emerge. When it comes to solving competition problems, however, between-group dynamics play a major role. They select *game changing norms* that affect the payoff related to the available strategies through punishment or reward to solve free-rider problems which create better equilibria than the ones originally available. It means that social arrangements with norms alter payoff matrices to ensure that self-interested strategies align with group interests, without requiring self-sacrifice. They are shaped by interactions between individuals and between groups, the latter selecting efficient equilibria and the former leading to salient ones. Sanctions are imposed to solve competition problems.

Gintis models norm enforcement and stability through *replicator dynamics* and public goods games. Suppose is the share of individuals using strategy (e.g., cooperating, defecting, punishing). Let be the fitness (expected payoff) of strategy . The replicator equation is:

where is the population average fitness. A norm is stable when the strategy it encodes becomes evolutionarily stable (resists invasion by mutants) due to its *adaptive advantage in group-level performance*.

What makes norms distinctive in Gintis’s account is the incorporation of *strong reciprocity*, a behavioral trait characterized by *cooperation with others and punishment of non-cooperators*, even at personal cost. Strong reciprocity is empirically observed in cross-cultural behavioral experiments like ultimatum, trust, and public goods games and contradicts the predictions of purely self-interested models (**gintis2005?**). Gintis treats this trait not as an anomaly but as an *evolutionary stable behavioral phenotype*, sustained through norm-based socialization and group selection.

A central and innovative concept in Gintis’s theory of social norms is the idea that norms transform not just preferences but the *structure of the strategic interaction itself*, by modifying agents’ *subjective representations of payoffs and actions*. This transformation is encoded in what he calls a *belief matrix*, a mapping of how agents perceive and evaluate their strategic options based on the presence of social norms (Gintis 2009a, ch. 12; **gintis2003?**).

In classical game theory, a game is defined by:

* A set of players ,
* A set of strategies for each player ,
* A utility function assigning payoffs.

Gintis argues that this framework is incomplete for modeling *norm-governed behavior*, because it assumes that agents evaluate strategies based on static utility functions. However, *norms induce endogenous changes* in the utility functions themselves, via socially learned expectations, emotions like guilt or shame, and reputational incentives. These are captured through a *modified payoff function*:

where is the *norm-adjusted utility*, and encodes *normative valuations* of strategy profile . The function depends on agent ’s *beliefs* about what is expected, appropriate, or punishable—thus forming part of a *belief matrix*.

The *belief matrix* is not merely a list of beliefs but a *second-order cognitive structure*: it encodes how players *transform the base game* into a normatively laden one. For example, in a Prisoner’s Dilemma, if both players believe that mutual defection is morally wrong and likely to incur reputational loss, their payoff matrix is *endogenously transformed* into a coordination game or even a Stag Hunt, depending on the intensity of normative beliefs. This resembles Crawford’s and Ostrom’s cooperation games -parametrized with incurred sanctions which I mentioned earlier.

To formalize this, let be the original payoff matrix, and be the *belief matrix* that maps social expectations, punishments, and rewards into numerical modifiers. Then:

where is the *norm-governed game* actually perceived and enacted by players.

This idea closely parallels Gintis’s general theory of *“strongly endogenous games”* (Gintis 2009a, 187–89), in which preferences and payoffs are not fixed but shaped by cultural and institutional context. Here, *social norms act as priors or filters* that reshape the game. The belief matrix may itself evolve over time, via cultural transmission, education, or feedback from repeated play.

Gintis thus provides a mechanism for the *cognitive embedding of norms* in strategic behavior, bridging the rationalist structure of game theory with *evolutionary and cultural psychology*. This resembles Bicchieri’ notion of cognitive schemata and hints on its mechanism. Gintis’s approach contrasts sharply with static or exogenous models of norms like Lewis’s conventions, and aligns Gintis with *constructivist* and dynamic modeling traditions in behavioral economics.

Gintis and Young (Young 1998) share an interest in the emergence and stability of social norms. Young explains norm stability via stochastic evolutionary dynamics and local interaction, using resistance trees and Markov chains to model convergence to norms. Gintis, by contrast, provides a biocultural account in which norms co-evolve with cognition, social learning, and enforcement institutions. Moreover, while Young focuses on punishment as a strategy, Gintis integrates it as an evolved emotional mechanism, part of the human behavioral repertoire.

Gintis’s theory positions norms as culturally transmitted and biologically grounded mechanisms for sustaining large-scale cooperation. Unlike equilibrium or expectation-based theories, his model embeds norm-following in the coevolution of genes and culture, and explains persistence through multi-level selection. Norms, in his view, are:

* Emotionally regulated,
* Costly but group-beneficial,
* Transmitted via imitation and enforcement, and
* Fundamental to the evolution of human societies.

Morsky and Akçay (2019) offer a rigorous evolutionary game-theoretic model of emergence of social norms from random beliefs. Their approach conceptualizes norms as *sets of behavioral prescriptions coupled with environmental descriptions*, which coordinate behavior via CE. Morsky and Akçay discern:

* *Behavioral prescriptions:* The set of actions that a norm holder is expected to perform in given contexts
* *Environmental descriptions:* The agent’s beliefs about the expected behaviors of others, conditioned on exogenous environmental events.

Crucially, environmental events have *no intrinsic normative meaning or direct impact on payoffs*. Instead, their significance arises from the *correlation structure* generated by agents’ shared beliefs and coordinated behaviors. This captures how seemingly arbitrary or “irrelevant” environmental cues can become *focal points* for coordination, akin to conventions or superstitions that facilitate social order. This conceptualization aligns closely with our focus on how *ontic correlations* become *epistemic*.

In Morsky’s framework, social norms function as *endogenous correlating devices*: the shared beliefs about environmental signals effectively serve as coordination signals that synchronize agents’ behavior. This mechanism explains how norms can be *self-enforcing* and *evolutionarily stable*, as agents find it rational to conform to expectations given their beliefs about others’ behavior.

A key insight of Morsky’s model is that social norms can emerge from *random accumulations of beliefs or superstitions* about environmental events that initially have no normative significance. Through repeated interactions and social learning, these beliefs become *correlated and self-reinforcing*, giving rise to structured norms that coordinate behavior effectively.

Morsky’s model emphasizes that *exogenous environmental events*, though initially meaningless, become meaningful through their *statistical correlation with agents’ behaviors*. These signals act as *coordination devices*, enabling agents to align expectations and actions. This insight aligns with the concept of *niche construction*: agents modify or leverage environmental features to create *shared external scaffolds* that reduce uncertainty and facilitate coordination.

Morsky and Akçay develop a formal model that captures how social norms operate in structured interaction environments. The world of events is modeled as an *undirected graph*:

* *Vertices* : Represent *privately observed events* or *contexts* in which interactions (games) occur
* *Edges* : Represent *interactions* between two players via a game. Each game corresponds to one edge, and each player in that game is assigned to one of the edge’s vertices (randomly and privately).

This event graph essentially defines a *joint distribution over events*, similar to the correlated signal structure in CE, but with a more flexible structure allowing *aggregation and labeling*.

Each edge is sampled with probability , where is a *symmetric weighted adjacency matrix*. So edge selection reflects the frequency or importance of particular interactions. When an edge is chosen, each player is *randomly assigned* one of its two vertices. If the edge is a self-loop, both players are assigned the *same* vertex, but still observe it *privately*.

To model limited observational resolution, the model introduces a set of labels , and defines a label matrix such that:

This formalism means that players cannot distinguish between vertices with the same label, thus introducing *ambiguity* into what they observe during interactions.

A *social norm* in this model is defined as a triplet of matrices:

* *Label matrix* : Describes how vertices are grouped into labels
* *Prescriptive behavior matrix* : Specifies what strategy a player should play when they observe a given label. That is, if strategy is prescribed for label
* *Descriptive behavior matrix* : Represents a player’s *beliefs* about what strategies opponents will play when they see a given label.

If the label set is as fine-grained as the vertex set (i.e., ), players can be prescribed specific strategies for each unique context. If , players generalize their behavior across indistinguishable situations.

When a norm lacks a prescription for a particular label, players revert to a *default behavior*, typically a *symmetric Nash equilibrium* strategy.

Players follow their own norm regardless of what their opponent does. Because of differing prescriptive matrices and , the *actual payoff* a player receives can deviate from expectations based on the descriptive matrix . Norms evolve not through individual introspection, but via *imitation dynamics*: players imitate those whose norms yield higher fitness.

The expected payoff to a player using norm with prescriptive matrix against an opponent using norm with prescriptive matrix is:

where:

* : A payoff matrix defining payoffs for each strategy pairing
* : Label matrices of the focal player and opponent
* : Their prescriptive strategy matrices
* : The adjacency matrix of the event graph.

This formula accounts for:

* The distribution over events (via )
* How those events are labeled (via and )
* What strategies are prescribed in response to each label (via and )
* And the game payoffs (via )

Importantly, (descriptive matrix) does not directly affect payoffs but only shapes whether players follow or deviate.

Morsky and Akçay’s framework allows us to mathematically represent social norms in structured populations as label-based prescriptions over event graphs. By modeling interactions probabilistically, assigning context labels, and embedding norm-following in matrix operations, the framework supports the study of norm dynamics, strategy evolution, and the role of ambiguity and coordination in norm emergence.

The authors apply their model to canonical games such as “Chicken” which is a non-evolutionary version of HDB. They show how norms can resolve coordination failures by correlating actions on environmental signals. This illustrates how environmental correlation structures and norm dynamics influence the resolution of social dilemmas.

Although this model is pretty close to what I want to achieve, Morsky’s agents are already initialized with *random private beliefs* over actions or strategies. While these beliefs may be uninformed or uncoordinated, they are *already epistemic*: they exist “inside agents’ heads” and represent subjective probabilities. By contrast, my concern is how *shared beliefs emerge from ontic regularities* — that is, from mere statistical patterns in behavior or environment like one agent dropping a token or choosing a location consistently, without presupposing internal belief states.

In addition, in Morsky’s model, the structure of the game is fixed. The CE emerges from updating beliefs based on observed behavior, but *the environment remains inert* — it does not reflect or encode behavioral history. I am seeking to show how agents *actively construct their informational environment* with markers, cues or patterns, which then feeds back into behavior. This dynamic coupling between agent and world is absent in Morsky’s approach.

Lastly, in Morsky’s model, CE emerge as stable mappings from beliefs to actions. But agents:

* Do not form *decoupled rules* like “play strategy X regardless of what others are doing because X is the rule now”
* Do not operate on *symbolic or generalized representations*
* Do not treat deviations as \*violations of expected rationality.

All in all, Morsky’s model is useful for explaining how coordination and equilibrium can emerge *from uncoordinated beliefs* via learning. But I seek to explain something else: *how shared belief and normativity themselves emerge* from mere observed regularities in behavior and the environment, and how agents construct a *(proto)-normative symbolic order* through informational niche construction. Hence, while Morsky’s work is adjacent, it is *insufficient* for my aim.

Overall, the dialogue between Lewisian conventions and social norms reflects a trajectory from descriptive equilibria towards normatively rich social arrangements. While Lewisian theory captures the elegance of coordination through common knowledge, the subsequent scholarship spanning Ullmann‑Margalit’s norm genesis, Epstein’s simulation models, Young’s evolutionary stability, Bicchieri’s conditional preferences, and Gintis’s cultural evolution reveals the layered complexity of societal regulation. Together, these perspectives elucidate how conventions can acquire prescriptive force, how norms are institutionalized and enforced, and how the interplay of rational choice, psychology, and evolutionary dynamics shapes the fabric of social life.

What has emerged as a pattern through the mentioned works is the continual relationship between conventions and norms. To make it more vivid, O’Connor (2019) draws two important distinctions:

* conventions and social norms,
* functional and arbitrary conventions.

First distinction means that, as we have seen, not all behavioral regularities have normative force in the sense of added layer of normativity. Friends having a convention of meeting each Friday evening at a bar, would barely be upset if someone has not showed up, because they do not possess a normative expectation or any other added layer of normativity. On the contrary, if two cars are driving on the same side of the road towards each other, the drivers are forced to swerve, for otherwise they might crash. They ought to swerve, for not only might one of them be fined but they might cause an accident. As Bicchieri (2005) points out, conventions are different from social norms in the relationship between self-interest and common interest. They coincide in the former and do not necessarily coincide in the latter. In the case of friends at a bar, there is no or little tension between self-interest and common interest, while in the case of driving cars there is. O’Connor stresses that conventions and norms are the *poles of a continuum* along which the former acquire normative force.

The second distinction O’Connor draws concerns the arbitrary and historically contingent nature of conventions that they “might have been otherwise”, which we will now review.

### Overestimation of arbitrariness

Overestimation of arbitrariness is another area of criticism of Lewis’s theory of conventions. Arbitrariness means that there are several equilprobable conventions which might have been otherwise. According to Lewis, arbitrariness is one of the distinguishing aspects of conventions. However, as Gilbert (1992) points out, not all possible solutions to a coordination problem are equally profitable for players. In cases where one way of coordinating is more preferred than another, convention will not be that arbitrary. In other words, alternative conventions are logically justified, but pragmatically implausible as there is almost always a slight “preference” of one convention over the other due to different factors like historical accident and history of play. O’Connor (2019) studied emergence of unfair norms like gendered division of labor in a similar vein. And later scholars talked about this in terms of symmetry-breaking by stochastic events (Skyrms 2010a, 2010b) and salience of conventions amplified by the history of iteratively played coordination game (Korbak et al. 2021).

As I mentioned, O’Connor (2019) sees arbitrariness as a continuum between contingency and necessity, or conventionality and functionality of conventions. Signaling between vervet monkeys might well be modeled as a convention in the Lewisian sense of repeated behavioral patterns of solving coordination problems (Harms 2004; Skyrms 2010a). However, this convention is not historically contingent in the sense of several possible solutions being equally profitable as Lewis supposes and as Gilbert critiques, for there are evolutionary constraints breaking the symmetry between multiple equilibria. Agents might be (and most probably are) hardwired to following certain strategies in certain environmental conditions. This distinction, as O’Connor underlines, highlights some conventions as more functional and others as more arbitrary.

A similar line of criticism came from Burge (1975), who noted that Lewisian requirement for convention to involve mutual knowledge of *alternative regularities*, or practices that could replace existing ones if widely adopted, is too strict. Conventions might fix without agents’ knowledge of alternatives, Burge argued. He contended that conventions can stabilize with habit, custom or tradition, widely following Hume’s original argument, and that knowledge of alternative conventions is not needed. Conventions, as Burge put forward, are not governed by any biological, psychological or sociological law, they are historically accident. In addition, agents do not necessarily deliberate to “switch” from one convention to another. In terms of game theory, Lewis required that agents know the structure of the game with its multiple equilibria, whereas Burge’s notion does not. This leads to yet another point of criticism, overly intellectualist requirements for agents.

O’Connor (2021) has proposed an information-theoretic measure of arbitrariness applicable to both human and “animal” conventions like alarm calls. It helps to break the distinction of functional and arbitrary conventions that she herself pinpoints (O’Connor 2019). As she is most interested in the emergence of cultural traits like gendered division of labor, she says that most cultural traits are both functional and arbitrary, or contingent, for they “might have been otherwise”.

O’Connor notes that Lewis’s notion of convention emphasizes arbitrariness, for a coordination game has at least two *proper coordination equilibria*. It means that either of them might have been established equiprobably. Gilbert (1992) critiqued this notion of arbitrariness and suggested that some equilibria are more favorable than others. In line with this, Simons and Zollman (2019) illustrated the distinction of functional and arbitrary conventions by putting it along three dimensions:

1. Payoff difference — some coordination equilibria have bigger payoffs than others;
2. Likelihood of emergence — some conventions are more likely to emerge than others;
3. Stability — once these conventions have emerged, they are unlikely to be deviated from.

O’Connor proposed to layer these dimensions onto evolutionary models, namely those of replicator dynamics. It allows for specifying what Simons and Zollman mean by the likelihood of emergence (2) and stability of conventions (3).

$$
\begin{array}{|c|c|c|}
\hline
& A & B \\ \hline
A & (1,1) & (0,0) \\ \hline
B & (0,0) & (x,x) \\ \hline
\end{array}
$$

A coordination game where B equilibrium is more favorable given

Modelled as replicator dynamics, the game has B equilibrium as more “natural” in the sense (1) — that of a higher payoff. The corollary of this is a larger basin of attraction, which represent the probability that each outcome evolves, given little information about the initial conditions of the population. Thus, B is more likely to emerge and is more “natural” in the sense (2).

To propose a measure of conventionality, O’Connor focuses on naturalness of conventions as their probability of emergence (2). As has been shown, different equilibria might evolve with different probability, and the sizes of basins indicate the amount of information we gain from examining an evolutionary process. The amount of information in this process is a measure of arbitrariness of a convention. It increases with greater uncertainty about what will evolve and decreases with less.

The relevant information-theoretic measure is Shannon entropy. It measures the average amount of information transmitted through a channel:

The amount of information gained from observing something is related to how much we learn or how surprised we are. It is calculated by summing the probabilities of the signals multiplied by their informational content , with equal to — the less probable a signal, the more information it carries. Overall, this weights the probability of each signal being sent by the amount of information it carries, giving a measure of average information in the channel.

If other conditions hold, a channel has higher entropy when signals are more equiprobable or there are more signals. This is as opposed to a biased channel, with one signal sent 99% of the time, which has an entropy of just . Therefore, more equiprobable signals and more signals lead to higher entropy. As probabilities represent the sizes of the basins of attraction, learning more from an evolutionary process increases the arbitrariness of an evolving phenomenon in question. Given more equilibria or if their basins of attraction are close to equal, the phenomenon is more arbitrary, contingent or “conventional”.

It is possible to measure the naturalness of a convention similar to Simons and Zollman (2019) using , which represents the informational value of a particular outcome. The closer the value of to , the more natural a convention is. This measure can be used whenever we have clear probabilities for different evolutionary outcomes. However, if there is uncertainty about the initial conditions of a population, it is possible to assign probabilities to different population starting places and not their basins of attraction, for they will not track the probability of emergence correctly. Another case is stochastic dynamics, where each starting point might lead to multiple equilibria depending on chance events. Here, can be defined as percentage of emergence of different equilibria in an iterated game. For example, if 10 individuals play the game in the matrix above with 10K times, it yields 6% chance of emergence of A, and 94% for B.

As O’Connor notes, one problem with a proposed conventionality measure is the source of probabilities. It means that inputs in an entropy equation can be selected quite arbitrarily depending on a case (ironically enough). As there are no actual chances, it can be hard to determine, whether a convention might have been otherwise. To address this worry, she underlines the *representation-dependence* of the measure. It means that we can specify what is probability in terms of a particular model or data set, for instance, basins of attraction, emergence over multiple rounds of simulation, equilibrium time and percentage of societies adopting a behaviour. This eliminates worries about probability and chance in the world. However, representations should be tailored to the intended explanatory goals. However, as representations have limitations, the proposed measure should not be taken as an absolute truth.

Another interesting approach to the notion of functionality of conventions is due to Harms (2004) whi specifically studies emergence of meaning in signaling games throughout evolutionary processes. According to him, any semantic convention, or “rule”, might be considered a “function-stabilizing mechanism”, meaning that a certain convention exists, because it has been selected to meet a certain evolutionary pressure. It helps to coordinate the behavior of organisms or different parts of an organism to perform an evolutionary adapted biological function. Rules are sets of maps from conditions to processes. They say what to happen next given a state of the world. Rules for evolutionary adapted traits (AT) might be expressed as:

$$R\_{AT}={\{\langle c\_{i,}p\_{i}\rangle} \mid AT \space sel \space p\_{i}\space in \space c\_i\}$$

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 in the conditions . (Harms 2004, 203).

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” (Seyfarth and Cheney 1990; Baraghith 2019). Harms calls this “primitive content” that has both indicative and imperative functions (Harms 2004, 189). Millikan calls it “pushmi-pullyu” representations and notes that purely descriptive and directive representations require a more advanced cognitive processes (Ruth Garrett Millikan 2005, 166).

Evolutionary development of primitive content leads to the divergence of its descriptive and directive functions due to advanced cognitive capacities. As Harms suggested, it introduces a stabilizing, or regulatory mechanism that works “atop” of conventions as rules for adaptive traits and guides behavior in case of failure of . It employs a corrective signal to “enforce” the initial convention when a signal is not sent in the presence of a world state it was selected for:

$$R\_{SM}={\{\langle \sigma\_{i} \wedge \neg m\_{j} \space where \space \langle \sigma\_{i,}m\_{j}\rangle} \in R\_{AT} \rangle\mid SM \space sel \space cs \space when \space (\sigma\_{i} \wedge \neg m\_{j})\}$$

We can already see parallels with Gintis in terms of “second-order cognitive structure” implied by metacognitive tracking of a corrective mechanism. The rule for a stabilizing mechanism is a set of ordered pairs consisting of the failure of an adaptive trait in the first place and a corresponding corrective signal in the second place. If the adaptive trait fails, the stabilizing mechanism will detect this failure and send a corrective signal/action to restore it.[[9]](#footnote-43) This division is echoed in Millikan’s work as first-order and higher-order reproductive families (Ruth Garrett Millikan 1987, 23). According to it, conventions are first-order and stabilizing mechanisms are second-order reproductive families that serve the same goal of restoring a first-order proper function. Although quote different in nature and in used language, Harms’s contribution shows how conventions can be almost purely functional and at the same time adequately incorporate proto-normativity.

Overall, the problem of arbitrariness in Lewis’s conventions has generated fruitful responses, especially in the naturalistic camp. A tightly connected notion is that of cognitive demands and common knowledge.

### Epistemic overreach of common knowledge requirement

Common knowledge denotes an epistemic state within a group wherein a proposition *p* is known by all members, and each member knows that every other member knows *p*, recursively extending to an infinite level of iterated knowledge. This recursive nature differentiates it from mutual knowledge, which necessitates only that each individual knows *p*. Consequently, common knowledge represents an idealized, stringent condition profoundly impacting coordination and strategic interaction, prompting investigation into its feasibility and real-world relevance.

As Cubitt and Sugden (2003) underline, Lewis’s initial conception of common knowledge did not imply unconstrained cognitive capacity of idealized agents. As they put forward, a proposition *p* is common knowledge if a state of affairs *A* exists where everyone has a reason to believe *A* holds, *A* indicates to everyone that everyone has a reason to believe *A* holds, and *A* indicates to everyone that *p*. This definition generates an infinite chain of “reasons to believe” rather than an infinite chain of “knowledge” as justified true belief suggesting a more pragmatic approach towards achieving coordination. This approach acknowledges the limitations of human epistemic capabilities and focuses on the justification for beliefs about states of affairs and others’ beliefs about them rather than in absolute certainty on every level of iterated knowledge. Nevertheless, the majority of scholars interpret Lewisian conventions as computationally and cognitively demanding.

Gilbert (1992) criticized the infinite regress of Lewis’s common knowledge. She challenged the psychologically implausible requirement of infinite levels of iterated knowledge, arguing it is unnecessary for explaining social phenomena like collective belief and convention. Gilbert proposed a framework centered on joint commitment, asserting that social facts emerge from situations where individuals are collectively committed to intend or believe something as a unified body, rather than through an infinite chain of individual beliefs about others’ beliefs. This joint commitment involves a shared intention or belief held by a group as a collective entity, irrespective of individual members’ personal convictions—for instance, a group’s shared commitment despite private doubts. This approach provides a means to understand shared social states and collective actions, generating shared obligations and expectations that drive behavior and shape attitudes, thereby avoiding the demanding epistemic requirements of common knowledge.

(**bicchieri1993?**) argued that real-world agents operate under bounded rationality, which is more psychologically plausible. Individuals possess finite processing capacity and memory, which makes an infinite regress of knowledge untenable. Bicchieri investigated how agents form beliefs and expectations about others’ actions in coordination games, emphasizing mutual expectations and the potential for coordination through learning and repeated interactions, even without full common knowledge. She highlighted the role of *social norms*, proposing that they function through conditional preferences – individuals preferring to conform if they expect others to do so – and normative expectations, which are beliefs about what others believe one ought to do. This allows coordination to emerge and persist through observation, belief updating, and conformity, irrespective of the norm’s common knowledge status.

(**heifetz1999?**) underscored the limitations of the common knowledge assumption in dynamic settings and games with temporal imprecision where communication is not instantaneous or unreliable. The coordinated attack problem when two parties agree to attack at the same time exemplifies how the absence of guaranteed, instantaneous communication can preclude the establishment of common knowledge, leading to suboptimal outcomes. Researchers have investigated alternative, weaker notions like finite levels of mutual knowledge or common belief to account for imperfections in real-world information and bounded rationality, offering potentially more accurate models of coordination and cooperation.

One of the more radical criticisms of the common knowledge requirement comes from evolutionary game theory, a branch of game theory pioneered by Maynard Smith (1982) which assumes natural selection and evolutionary dynamics as a source of solutions for strategic games instead of rationality of self-interested actors with complete information. These criticisms doubt the necessity of common knowledge for conventions.

For example, Binmore (2008) challenged the infinite levels of common knowledge posited by Lewis, arguing that agents only require first-order expectations regarding others’ behavior to converge on an equilibrium. This perspective emphasizes accurate prediction of actions as a critical element for coordination, with rational players responding accordingly. Binmore’s evolutionary approach highlighted cultural evolution’s role in shaping these common understandings and norms, suggesting societies develop and transmit effective coordination strategies over time based on promoting social stability, a dynamic process which refines coordination strategies rather than a static, pre-existing condition of full common knowledge. He also noted that Lewis’s analysis of conventions confines its usage to small-scale societies as it implies observing public events being observed by another party. And this is not realistic in larger populations. Binmore suggested that conventions do not generally require common knowledge overall and can be established in evolutionary environments with only one level of reasoning instead of infinite hierarchy of beliefs. He also notes that everyday conventions mostly operate via automatic behavior and low-level mutual expectations.

Guala (2020) put forward a similar argument about “belief-less” coordination where most everyday conventions do not require iterated beliefs and hence cognitive capacities for meta-representation. Means-ends rationality and cheap heuristics are said to be sufficient[[10]](#footnote-45).

Schiffer (1972) conceptualized common knowledge as an *infinite hierarchy of mutual knowledge*, defined recursively as follows:

* Everyone knows proposition
* Everyone knows that everyone knows
* This recursion continues for levels, where

Formally, common knowledge is the infinite conjunction:

where denotes the -th level of mutual knowledge. This approach emphasizes *communication* as the mechanism to elevate mutual knowledge into common knowledge. For example, if two children both see mud on each other’s faces, they initially have mutual knowledge as both know there’s at least one muddy face. Only after public announcements do they iteratively reach mutual knowledge₂, mutual knowledge₃, etc., converging toward common knowledge.

Aumann (1976) formalized common knowledge using *information partitions* over possible states of the world in his “agreement theorem”. If is the set of possible states, and represents the information partition of agent , common knowledge is an event at state if the cell of the *meet* of the agents’ partitions containing is a subset of .

A key result is Aumann’s “agreement theorem”: If two rational agents with common priors have common knowledge of their posterior probabilities about an event , their probabilities must coincide. Formally, if for agents , then . This framework avoids explicit infinite regress by defining common knowledge as a fixed point in the agents’ information partitions.

Schiffer’s model aligns with intuitive examples like the muddy children one, where announcements progressively deepen mutual knowledge. Aumann’s approach, by contrast, provides a mathematical foundation for strategic interactions, showing how common knowledge enforces consensus among Bayesian agents. While Schiffer’s hierarchy is often seen as philosophically intuitive, Aumann’s formalism has been more widely adopted in economics and computer science due to its technical precision. Both, however, agree that common knowledge transcends mere mutual understanding, requiring agents to account for *each other’s epistemic states* in a systematic way.

The relationship between Lewis’s notion and these subsequent formalizations remains debated. Cubitt and Sugden (2003) suggested a fundamental distinction between Lewis’s perspective and later theories, while Vanderschraaf (1998) and Sillari (**sillari2005?**) argued for continuity between them.

For Lewis, common knowledge is a *social construct* tied to conventions. Agents coordinate because they recognize a shared basis for action, not because they explicitly compute infinite epistemic hierarchies. Schiffer formalizes common knowledge as an infinite tower of mutual knowledge . Aumann models it via information partitions, common knowledge as the meet of agents’ partitions in . Both frameworks presuppose agents with unbounded logical capacity to process infinite iterations or partition intersections which contrasts with Lewis’s emphasis on pragmatic nature of coordination.

To substantiate the sharp distinction between Lewis’s and Shiffman/Aumann notions of common knowledge, Cubitt and Sugden (2003) highlight five key areas of divergence:

* *Epistemic and practical foundations*:
  + Lewis’s common knowledge is action-oriented and rooted in conventions (e.g., stopping at red lights)
  + Schiffer/Aumann treat it as a logical property of agents’ knowledge structures, decoupled from real-world coordination
* *Role of salience*:
  + Lewis requires salient cues (e.g., focal points) to bootstrap common knowledge
  + Schiffer/Aumann exclude salience, relying instead on axiomatic mutual reasoning
* *Bounded and unbounded rationality*:
  + Lewis’s agents operate under bounded rationality—common knowledge emerges from practical precedents, not recursive deductions
  + Schiffer/Aumann assume agents can process infinite hierarchies or partition operations, implying unbounded cognitive capacity
* *Necessity of publicity*:
  + Lewis emphasizes public events (e.g., a fire alarm) as triggers for common knowledge
  + Schiffer/Aumann reduce publicity to abstract logical or set-theoretic constructs (e.g., public announcements as partition refinements)

Cubitt and Sugden contend that Schiffer/Aumann’s formalisms fail to capture the pragmatic nature of common knowledge which involves culture and history in Lewis’s work. This disconnect, according to Cubitt and Sugden, renders Schiffer/Aumann models ill-suited for explaining real-world scenarios Lewis sought to address, such as language conventions or social norms. Their critique underscores that Lewis’s common knowledge is *procedural* and rooted in coordination practices, while Schiffer/Aumann’s is *declarative*, being a static property of knowledge structures. This philosophical and operational gap explains why formal models struggle to replicate the dynamism of Lewis’s convention-based framework.

Peter Vanderschraaf and Sillari (2023) argue for *continuity* between Lewis’s convention-based notion of common knowledge and the formalizations of Schiffer/Aumann, positing that these frameworks share foundational goals and can be reconciled through careful interpretation. Their analysis emphasizes complementary perspectives rather than irreconcilable differences.

To start, both accounts (Lewis’s and Shiffer/Aumann’s) share a goal of explaining coordination. Vanderschraaf and Sillari contend that Schiffer/Aumann’s models *operationalize* the abstract conditions Lewis presupposed in his conventions. For example, Aumann’s partitions can encode the “publicity” of events central to Lewis’s conventions, while Schiffer’s hierarchy mirrors the iterative reasoning agents use to infer shared expectations.

Lewis’s emphasis on *public events* like fire alarm as triggers for common knowledge aligns with Schiffer/Aumann’s focus on *public announcements* or *partition refinements*:

* In Aumann’s model, a public event corresponds to a partition cell known to all agents, which becomes common knowledge after observation
* Schiffer’s hierarchy similarly requires public communication to escalate mutual knowledge.

This bridges Lewis’s pragmatic notion of salience like a red traffic light with formal models’ requirement of shared information structures.

Next, Vanderschraaf and Sillari argue that Schiffer’s infinite hierarchy and Aumann’s partitions can accommodate *bounded rationality* in practice:

* Agents in real-world scenarios need not compute infinite recursions explicitly. Instead, established conventions act as *focal points* that *approximate* common knowledge after a few iterations[[11]](#footnote-46)
* For example, drivers stopping at a red light rely on a convention that implicitly assumes mutual knowledge up to a pragmatically sufficient level like or , bypassing infinite regress. This is called *fixed point definition*, because it “fixes” a common knowledge as a point in probability space (Cedric Paternotte 2011). This interpretation aligns closely with cognitive realism of *relevance theory* (Wilson and Sperber 1999; Sperber and Wilson 2002), where individuals arrive at mutual understanding having found the minimally interpretable meaning of a signal which aligns with context. It is also sometimes called *fast and frugal* heuristics, which optimizes cognitive resources by using simple rules like “stop searching if found minimally adequate meaning” (Gigerenzer and Todd 1999; Gigerenzer and Selten 2002). This aligns with experimental evidence using “cognitive hierarchy theory” showing that humans rarely reason beyond 2–3 levels of mutual knowledge in coordination games (Bardsley et al. 2010).

As Vanderschraaf and Sillari continue, Lewis’s conventions can be mapped to *fixed points* in Aumann’s partition framework:

* A convention corresponds to a *stable equilibrium* where agents’ knowledge partitions align with shared social rules
* Similarly, Schiffer’s hierarchy converges to a fixed point as mutual knowledge deepens through repeated interactions.

Thus, conventions are not merely cultural artifacts but *emergent properties* of epistemic structures formalized by Schiffer/Aumann.

Vanderschraaf and Sillari reinterpret Lewis’s “salience” as a mechanism for *partition refinement* or *hierarchy truncation*:

* A salient event publicly signals a shared context, allowing agents to coordinate without infinite reasoning
* This can be modeled in Aumann’s framework by restricting the state space to a subset where salience is common knowledge.

Both Lewis’s and Schiffer/Aumann frameworks implicitly treat common knowledge as a *dynamic process* rather than a static state.

* *Lewis*: Conventions evolve through precedent and reinforcement
* *Schiffer/Aumann*: Mutual knowledge escalates via announcements (Schiffer) or Bayesian updating (Aumann).

This shared emphasis on iterative learning undermines Cubitt and Sugden’s “discontinuity” claim, as both traditions acknowledge that common knowledge is *constructed* through interaction.

Overall, Vanderschraaf and Sillari’s continuity thesis hinges on three claims:

1. Schiffer/Aumann’s formalisms generalize Lewis’s conventions by specifying their epistemic preconditions
2. Bounded rationality in practice truncates infinite hierarchies to match Lewis’s emphasis on pragmatic coordination
3. Salience and publicity act as bridges between abstract formal models and real-world conventions.

By framing Lewis’s work as a *procedural instantiation* of Schiffer/Aumann’s declarative structures, they argue that the frameworks are complementary, not contradictory.

Overall, Lewis’s notion of common knowledge ignited ongoing debates in philosophy, cognitive science and logic regarding its nature. Throughout the thesis, I will stick with Aumann’s notion of common knowledge as information partition, because our main framework for studying conventions will be Vanderschraaf’s. Of course, there are other accounts of common knowledge with subtle aspects of it, which are not directly relevant to the present thesis (Cedric Paternotte 2011; Monderer and Samet 1989; Morris 1999; Bonnay and Egré 2009).

### Connection between conventions and coordination problems

Some scholars argue that conventions are not necessary for solving coordination problems, undermining Lewis’s theory. Sugden (2005) and P. Vanderschraaf (1998) argue that conventions need not necessarily be solutions for coordination problems. For instance, fashion or property conventions are not like this[[12]](#footnote-48). Both of them have developed generalized accounts which do not require conventions to solve coordination problems. (**davis2003?**), Marmor (1996, 2009), (**miller2001?**), Sugden (**sugden1986?**; **sugden2004?**) have argued that conventions need not be coordination equilibria.

Sugden (2005) posits that conventions arise from behavioral patterns generating *mutual advantage*, independent of explicit coordination, thus rejecting Lewis’s focus on purely coordinating problems. Drawing on Hayek and Hume, he emphasizes spontaneous order of conventions which challenges the primacy of “constitutive” pre-established rules like law in governing social interactions. He argues that conventions emerge when patterns of behavior yield benefits *for all participants*, even in competitive or asymmetric situations. Unlike traditional game-theoretic models that focus on Nash equilibria, Sugden’s framework accommodates scenarios where *no clear equilibrium exists* which renders Lewis-style coordination problems too restrictive.

Sugden introduces the concept of *team reasoning*, where individuals act on collective goals rather than individual incentives, akin to Gilbert’s joint commitment and collective intentionality but without endorsing a “plural subject” ontology. Fashion conventions emerge through independent adoption of trends perceived as advantageous for social signaling. This framework elucidates conventions in competitive scenarios lacking coordination equilibria, exemplified by property rights systems governed by historical precedent rather than coordinated agreement.

As (**davis2003?**) and Marmor (2009) note, people follow trends for social distinction rather than coordination, yet these patterns become conventional through repeated adoption.

Marmor (2009) challenges Lewis’s emphasis on coordination problems, arguing instead for an analysis grounded in actual games like chess, distinct from the theoretical “games” favored by game theorists. His main argument is that there are deeper conventions like truth-telling which make Lewis-style coordination possible. He outlines three conditions for a rule to be considered conventional:

1. A population normally follows rule in circumstances
2. There is a reason for members of to follow in circumstances
3. There exists at least one alternative rule , such that if members of had followed instead of , would still have been a sufficient reason for following , partly because was generally followed instead of . Rules and are mutually exclusive in the given circumstances.

Marmor draws two distinctions:

* coordination / constitutive conventions
* “deep” / “surface” conventions

Coordination conventions solve Lewis-style problems like driving sides by aligning actions for mutual benefit and depend on shared expectations and mutual compliance. Constitutive conventions create social practices or classes thereof like chess rules which constitute the game of chess itself. Marmor argues that constitutive conventions emerge as responses to complex social needs and are foundational to many practices, including legal systems. Unlike coordination conventions, they do not depend on mutual expectations but instead define the ontology of the practice. Deep conventions are foundational norms that underpin social practices and are less amenable to change. For example, truth-telling is a deep convention necessary for effective communication. In its turn, surface conventions are more specific instantiations of deep conventions and vary across contexts. For instance, particular linguistic rules like grammar are surface conventions based on deeper norms like truth-telling.

Ruth Garrett Millikan (2005) presents a radically biological perspective on conventions, diverging significantly from economic and sociological approaches. Her core argument posits that a convention is fundamentally a behavior pattern sustained within a population through the mechanism of replicated precedent. Notably, Millikan rejects the prevailing tradition, exemplified by Hume and Lewis, which attributes social order to the rational decisions of individual agents. She explicitly denies any role for rationality in convention maintenance, asserting that a society upholding a convention solely through unreflective conformity would fulfill her definition. While Burgé similarly emphasizes factors beyond enlightened self-interest including inertia, superstition, and ignorance, Millikan’s position is more extreme, entirely excluding any rational underpinning for convention stability.

For Millikan, conventions persist through replication adjusted according to the weight of precedent, where current patterns derive from prior instances. They are arbitrary and contingent as their stability is dictated neither by optimal design nor conformity of the majority. Instead, it is influenced by its effective functional performance which might have been achieved with other patterns and does not require conscious adherence to rules. Millikan’s approach characterizes conventions as descriptive regularities which are emergent, stabilized patterns replicated through unconscious imitation, allowing for flexible adaptation without rigid definitions or universal agreement. For example, language speakers do not consciously follow a rule when calling a book a “book”. Instead, they simply replicate the behavior they have observed and linguistic conventions can be disobeyed without incurring sanctions, unlike rules in a normative sense. This contrasts with Lewis’s high-demanding view of mutual expectations, common knowledge and inherent normativity. However, it resembles Guala’s notion of “belief-less” coordination.

Millikan distinguished three types of coordination:

* *blind coordination*, where participants act without knowledge of each other’s actions (e.g., traffic systems in Lewisian examples)
* *half-blind coordination*, where one party anticipates the other’s behavior based on precedent (e.g., linguistic communication)
* *open coordination*, where both parties fully anticipate each other’s actions.

Linguistic conventions predominantly fall into half-blind coordination.

Millikan’s biological perspective frames conventions as analogous to evolutionary processes:

* Just as genes propagate based on their fitness, cultural conventions proliferate because they serve useful functions for individuals or groups
* The “proper function” of a convention is its capacity to achieve specific outcomes (e.g., facilitating communication) effectively over time.

The notion of function will be important later as it is used in contemporary theories of social institutions as strategic equilibria (Guala and Hindriks 2015) which try to smuggle biological functions and generate controversy over the very notion and its relation to convention.

This concludes the overview of criticisms of Lewis’s theoy of conventions and broader problems it generated. In the next section, we will look into the important refinements of Lewis’s theory which will be of central importance in the next chapters, when we will discuss game-theoretic approaches to social ontology and the problem of ontic reference it generates.

## Extensions and refinements

Lewis’s theory of conventions became a starting point for formal research on conventions and later scholars refined his theory, sometimes to an unrecognizable extent. There are many refinements, but we will consider only most important for the topic of emergence of social institutions from animal conventions. We will cover equilibria concept refinements by Vanderschraaf and Skyrms and the notion of salience in its relation to arbitrariness/functionality of conventions. As we will see, all these aspects will come together in shaping the notion of naturalistic account of social institution in the next chapter.

As I have mentioned in the previous section, imprecise equilibrium concept was among the popular criticisms of Lewis’s theory, and this component has been actively worked and elaborated on. Two notable reformulations of conventions are as *correlated equilibria* (CE) and *evolutionary stable strategies* (ESS). We start by studying them.

### Vanderschraaf’s *inductive deliberation* as a source of salience

Vanderschraaf (1995, 2001; 1998) redefined social conventions as CE through *inductive learning*, positioning conventions as foundational to achieving justice as mutual advantage. He formalized the notion of salience (or focal points) as information partitions and employed the *Dirichlet rule*[[13]](#footnote-51) to show how agents sequentially update their beliefs about others’ strategies to gradually arrive at an equilibrium *endogenously*, without explicit external signal.

Lewis considered a coordination equilibrium a convention if the players have common knowledge of mutual expectations. Vanderschraaf calls this mutual expectation criterion (MEC). Each agent has a decisive reason to conform to their part of the convention, expecting the other agents to do likewise. Lewis stated 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.

According to Vanderschraaf, a convention constitutes a strategy profile where each agent maximizes expected utility such that for all alternative strategies , ensuring stability against unilateral deviations.

The formation of conventions operates not through cognitively expensive rational deliberation, but through relatively cheap *inductive learning* mechanisms. Agents employ *Dirichlet dynamics* to update beliefs about opponents’ strategies. This updating process describes how agents repeatedly revise their beliefs by incorporating new observations of others’ behavior. A *deliberational equilibrium* is then defined as a fixed point of this learning dynamic, where agents’ beliefs stabilize. The stabilized joint beliefs and strategies that emerge from this iterative updating correspond to what Vanderschraaf calls *endogenous correlated equilibrium* (ECE)[[14]](#footnote-52): a CE arising internally from the agents’ inductive learning and mutual belief revision, rather than from an external correlation device as it is usually presented in broader game theory literature[[15]](#footnote-53). Kôno (2008) has mathematically proven how ECE is possible and that distributions of ECE and exogenous CE are completely different. The Dirichlet dynamics responsible for arriving at ECE is modeled as follows:

where represents observed strategy profiles and denotes prior beliefs (Peter Vanderschraaf 2018). Repeated interactions lead to path-dependent emergence of focal points, particularly in bargaining scenarios. Two prominent conventions arise: equal division of goods () and egalitarian payoff distributions satisfying for all agents , where represents disagreement payoffs (Peter Vanderschraaf 1995).

An important part of Vanderschraaf’s theory of conventions is his contribution to moral philosophy and theory of justice. He grounded principles of justice in conventions that generate Pareto improvements[[16]](#footnote-54) over non-cooperative baselines. A just convention must satisfy for all agents , where denotes the baseline equilibrium (Peter Vanderschraaf 2018).

This requirement addresses the vulnerability objection to justice theories which fail to adequately protect the most vulnerable persons. It does so by ensuring that conventions benefit even the least advantaged participants, creating mutual advantages that stabilize social arrangements. The framework reconciles Humean conventionalism with game theory, demonstrating how justice emerges from repeated coordination problems rather than abstract moral principles[[17]](#footnote-55).

As can be seen, convention as CE allows for more “fair” coordination, even though no pure strategy equilibrium exists as we saw earlier with the “Battle of Sexes” game example. To reiterate, 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 both agents play their strategies with probability , yielding an expected payoff of 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 with “heads” and “tails”. The agents have a common information partition and the correlated strategy combination is denoted as a function with and . Husband has a higher expected payoff with this combination than any of the other strategies, so he will not deviate from it. The expected payoff for Husband is if the outcome is , and if it is .

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

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

For Vanderschraaf, a convention is a mapping of “states of the world” to strategy combinations of a noncooperative game (Peter Vanderschraaf 1995, 69):

A *game* is an ordered triple consisting of the following elements:

1. A finite set , called the *set of players*;
2. For each player , there is a finite set , called the *alternative pure strategies* for player . The Cartesian product is called the *pure strategy set* for the game ;
3. A map , called the *payoff function* on the pure strategy set. At each strategy combination , player ’s payoff is given by the -th component of the value of , that is, player ’s payoff , at is determined by

where projects onto its -th component.

As Vanderschraaf builds on Aumann’s model (1987), each player has a personal *information partition* of a probability space . Elementary events on are called *states of the world*. At each state , each player knows which element has occurred, but not which . represents ’s private information about the states of the world. While knows the opponent partitions, she does not know their content. A function defines a *exogenously correlated strategy -tuple*, such that at each state of the world , each player selects a strategy combination correlated with the state of the world . Thus, by playing , follows *Bayesian rationality* and maximizes expected payoff given private information and expectations regarding opponents.

In addition, given , , and the information partitions of as defined above, is a *correlated equilibrium* if and only if, for each ,

1. is an -measurable function, that is, for each , is constant for each , and
2. For each ,

where denotes expectation, ‘’ refer to the result of excluding the -th component from an -tuple. This holds for any -measurable function . The correlated equilibrium is *strict* if and only if the inequalities are all strict.

The measurability restriction on means that knows her strategy in each . This definition implies that players have common knowledge of the payoff structure, partitions of , and , 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 , 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 has a private information partition of , there is a partition of , namely the intersection , of the states of the world such that for each , all the agents will know which cell occurs. The agents’ expected utilities in the following Definition 3 are conditional on their common partition , 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 of the states of the world, which is the intersection of all their private partitions . This reflects that conventions depend on information available to all agents.

Given , and the partition of of events that are common knowledge among the players, a function is a convention if and only if for each , and for each is -measurable and

for each and for any -measurable function .

It means that if any player deviates from a convention , every player , including , 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 will play a strategy other than the one prescribed by , they may be tempted to deviate, resulting in a worse-off outcome for . Conversely, if all opponents are aware that will play her strategy at each state of the world , then they have a strong incentive to conform with convention , which gives an improved outcome.

Overall, Vanderschraaf’s contribution is formalization of salience, hence he uses the *common* information partition as a necessary restriction to make the definition of convention conform with Lewis’s spirit. The other question is how salience itself emerges. Lewis suggested 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 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 (Peter Vanderschraaf and Skyrms 1993), proposed *inductive deliberation* as a mechanism by which salience is being established. It *requires agents to be conditional* 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.

We see here Bayesian rationality, dynamical updating and capacity for recursive beliefs as features of a certain *cognitive architecture* of an agent, a characterization of their cognitive capacities which influence their behavior. As we will see later, this implicit notion of agent’s cognitive architecture will be important in the discussion of social ontology in the next chapter.

Another significant extension of Lewis’s theory is related to redefining conventions as ESS and is due mostly to Skyrms.

### Skyrms’s evolutionary approach to conventions

Skyrms integrated Lewis’s theory of conventions into an evolutionary framework. He showed how signaling conventions can emerge naturally with adaptive processes like evolution and learning in agents with limited cognitive sophistication which overcomes Lewis’s reliance on common knowledge (Skyrms 2010a).

Although Skyrms has almost established an entire research program with many followers (Huttegger 2007a, 2007b; O’Connor 2020; LaCroix 2020a; Franke and Wagner 2014) and we will take a closer look at his generalization of Lewis’s signaling models later in this section, I suggest he would not have done it without his earlier and less-known contribution to game theory which has to do with generalization of the ESS solution concept.

The ESS, or evolutionary stable strategy, being a foundational solution concept in evolutionary game theory formulated by J. Maynard Smith and Price (1973) is a strategy that, if adopted by majority of population, cannot be invaded by any mutant strategy. Crucially, this concept implies random matching[[18]](#footnote-57), where individuals are paired for strategic interactions independently of their types, such that the probability of encountering any strategy is only proportional to its overall population frequency. While this assumption simplifies analysis and yields elegant theoretical results, it limits the applicability of ESS to well-mixed populations and fails to capture the complexity of structured or socially embedded interactions.

Skyrms recognized that ESS does not generate *stable* strategies with non-random matching arising from mechanisms like kin selection, signaling systems, spatial or social structures. These correlations induce interactional dependencies increasing the probability of similar-strategy encounters. Such dependencies drastically alter the evolutionary dynamics and can stabilize strategies such as cooperation or signaling conventions that would be unstable or unsustainable under classical ESS assumptions (Skyrms 1994).

This led Skyrms to establishing “adaptive ratifiable strategy” as a generalization ESS that incorporates the endogenous structure of interactions, making it a more realistic predictor of evolutionary outcomes. A strategy is adaptive-ratifiable if it maximizes expected fitness when it is nearly fixed in the population, taking into account the conditional probabilities of interacting with other strategies. This concept ensures dynamic stability under replicator dynamics where correlation affects interaction frequencies (Skyrms 1994).

The notion of adaptive ratifiable strategy made another Skyrms’s concept possible. That of “correlated convention” (Skyrms 2014a), which is conventions as stable yet not necessary Pareto optimal behavioral patterns made possible due to interactional dependencies of any kind between agents. Skyrms explored many possibilities for such correlation like spatial interaction (J. Alexander and Skyrms 1999), social structure (Skyrms 2003), social networks (Skyrms and Pemantle 2004) and finally signaling systems (Skyrms 2010b). However, as we will see in the second chapter of the thesis, Skyrms’s “correlation” is different from Vanferschraaf’s.

Skyrms’s approach to conventions differs from Lewis’s in not relying on common knowledge and substituting it with evolutionary pressures which make conventions arise and persist. He showed that even simplest organisms like bacteria can arrive at signaling systems akin to Lewisian conventions with the aid of simple adaptive mechanisms like mutation-selection or reinforcement learning (RL) (Skyrms 2014a).

Skyrms explored various learning dynamics that enable signaling systems to emerge in populations. For example:

* *Simple RL* where agents adjust their strategies based on trial-and-error feedback from successful interactions. In a basic Lewis-Skyrms signaling game setup with 2 world states, 2 signals and 2 actions, senders and receivers begin with random dispositions and gradually reinforce successful pairings between states, signals, and actions.
* *Win-Stay/Lose-Shift dynamics* where agents establish conventions more rapidly than simple reinforcement learning. This dynamic involves sticking with successful strategies while shifting away from unsuccessful ones, enhancing convergence speed and stability.

Skyrms’s framework models conventions as stable equilibria of sender-receiver games that evolve via RL and evolutionary dynamics rather than rational deliberation. Formally, a signaling game involves:

* a set of states of the world
* a set of signals
* a set of acts .

The sender observes a state and chooses a signal to send. The receiver, upon receiving , chooses an action . The payoffs and for sender and receiver respectively depend on how well the receiver’s action matches the state. Unlike Lewis’s model, which assumes common knowledge of salience to coordinate on a unique equilibrium, Skyrms shows that conventions can emerge through adaptive processes even when initial behaviors are random and no focal points exist.

A central concept in Skyrms’ analysis is the informational content of signals, which he quantifies using information-theoretic measures. Given a prior probability distribution over states and a posterior distribution conditioned on a signal , denoted , the information conveyed by can be expressed as the vector of log-likelihood ratios:

where represents prior probabilities of states and denotes posterior probabilities conditioned on a signal . This formalization bridges Lewis’s conceptual framework with mathematical models of communication.

This measure captures how a signal updates the receiver’s conditional strategy choice given the state of the world, thereby guiding action selection (Skyrms 2010a).

Skyrms further explores signaling equilibria under conditions of partial alignment or conflict of interests between sender and receiver. In such cases, the equilibrium strategies may involve deceptive or partially informative signals. Formally, if a sender’s payoff function differs from the receiver’s , the equilibrium concept extends to signaling equilibria where strategies and satisfy mutual best responses:

where denotes the set of probability distributions over (**skyrms1996?**).

The evolutionary dynamics driving the emergence of conventions are often modeled through RL algorithms such as the Roth-Erev model (Erev and Roth 1998). Agents maintain propensities for choosing actions (signals or responses), which are updated iteratively according to received payoffs:

where is a learning rate and is the reward at time for action (Skyrms 2010a). Over repeated interactions, these learning dynamics lead to convergence on stable signaling conventions without requiring explicit coordination or rational foresight.

Transmission of information in signals and emphasis on *informational content* of a signal generated a heated debates in philosophy of biology critiquing Skyrms for the lack of causation (Shea, Godfrey-Smith, and Cao 2018; Godfrey-Smith 2020; Harms 2004) and over-reliance on statistical connection instead of functional one.

An interesting part of Skyrms’s extension of Lewis’s signaling game is its implicit reliance on epistemic language of “observing” states of the world and “interpreting” signals for “updating beliefs”. Although Skyrms utterly rejects any Bayesian interpretation of his signaling games (LaCroix 2020b), he is sometimes interpreted as a incurring epistemology to his agents, especially when his theory is discussed side-by-side with natural theories of mental content (Ruth Garrett Millikan 1987; Ruth Garrett Millikan and Millikan 2004; Baraghith 2019; Harms 2004): that senders “represent” world states and transmit this public representation to a receiver who then “interprets” it with its own mental states. Consider vervet monkeys’ alarm calls. They can easily be described as involving mental states of “representing” an eagle and sending a certain signal to fellows monkeys who “decode” that public representation and map it onto suitable action. While plausible and the case for most natural theories of mental content like Ruth Garrett Millikan and Millikan (2004), it is not the case for Skyrms.

Although the structure of Lewis-Skyrms game mirrors the flow of information in epistemic contexts (state-signal-action pairings) and it is tempting to treat senders and receivers as conditional, the Skyrmsian agents update their *behavioral dispositions* rather than beliefs as they do not possess any inference and can only adjust their mappings according to failure rates (Skyrms 2012).

Skyrms’s sender-receiver system is an *information channel* focusing on how effective codes (signal-meaning pairings) arise and stabilize, not on agents’ beliefs or intention. His signaling games are mechanistic as Maynard Smith’s, for they take into account only objective, or “ontic”, features of agents like strategy frequency across population or, in case of signaling game, *mappings* from state to signal and from signal to action in accordance to the rate of coordination failures. Compare Lewis-Skyrms game

$$
\begin{array}{ccccc}
World & \xrightarrow{state} & Sender & \xrightarrow{Message} & Receiver & \xrightarrow{act} & {} \\
\end{array} \\
$$

with Shannon’s information channel:

Martínez (2019) proposes a “channel-first” view on signaling games and argues that the central behavioral unit of Lewis-Skyrms games is not strategies, but the *encoding-decoding pair* which is similar to *mappings* from above.

In this framework, world states, signals and actions can be represented as *random variables* , and , each of which is a set of discrete units like states, messages and actions like together with a probability distribution over them. The same applies to messages and actions.

A sender observes the current state and transmits a signal – one of possible signals. The receiver detects this signal and chooses an action, , from a set of available actions. Both the signal sent and the action chosen are random variables.

The probabilities for the random variables are linked through the sender’s and receiver’s strategies which are a probability matrices of signals given world states of acts given signals respectively.

Sender and reciever probabilities in Skyrms (2010a)

As per criticisms of Skyrms’s approach to Lewisian signaling games, Martínez (2019) argues that Skyrms did not go far enough into information theory and allowed informational analysis only *after* sender and receiver adopted the strategies which does not explain *how they arrived* at them. Martinez suggests using Shannon’s rate-distortion function (Shannon 1948) to show minimum mutual information between states and acts with minimum rate of distortion. It allows him to recast payoffs as distortion indicators in the channel. Seen with this lens, a coordination game of signaling, even that involving deception, as an information channel looks more cooperative.

Overall, Skyrms’s extension of Lewis’s theory of conventions has dropped rationality requirements and introduced a more naturalistic account of signaling systems in a broader context. Crucially, it implies a minimal cognitive architecture (or a lack of it) drastically different from conditional agents of Vanderschraaf.

### Salience in coordination games: epistemic and natural

Although both Vanderschraaf’s and Skyrms’s refinements of Lewis’s theory might seem concerned only with its formal details, they essentially tackle a wider problem of *salience* or *focality* of strategies which is tightly connected with “common knowledge” problem: how and why some strategies (and profiles thereof) “stand out” in the face of other arbitrary arrangements for potential conventions. Vanderschraaf formalizes salience as information partition of agents’ joint probability distribution and Skyrms formalizes it as emergent property of interactional asymmetry of agents’ or environment’s properties in evolutionary rather than Bayesian “update”. Here we briefly review approaches to the problem of salience of strategies as they will be relevant in the next chapters.

We can distinguish between two types of salience: epistemic (Zachník 2021; **mehta2021?**) and natural (VanDrunen and Herrmann 2023). Broadly speaking, epistemic salience has to do with beliefs of agents, whereas natural salience arises from environmental features or cues to which agents attend.

For Lewis, as he wanted to build upon Schelling’s concept of “focal point”, salience was unquestionable, but its very notion generated problems as to why and how certain features of environment, history and culture become relevant as cues external to coordination games in the first place. Schelling pointed out that to look for and *recognize patterns* might be an essential feature of human nature (Schelling 1980, 104). Observable patterns, he argued, presuppose their observability, hence it is rational to expect an opponent to behave according to such patterns and thus deduce intentions of another opponent (Schelling 1980, 104). A similar notion is found in “Relevance theory” (Wilson and Sperber 1999; Sperber 1996) in cognitive science, where an signal presupposes its optimal relevance, meaning that a signal is “articulated enough” across any sensory modality it is automatically worth noticing, implying that it is a universal feature of human and possible animal cognition. Consider a game-theoretic example from Guala and Hindriks (2015).

Dinka and Nuer tribes graze their cattle on two sides of the river Sobat. After the river dried up, its physical barrier disappeared, creating a Hawk-Dove conflict[[19]](#footnote-59) over grazing lands where both tribes risked costly clashes if they chose the same territory. The dry riverbed, however, remained a salient and mutually recognized landmark, providing a natural coordination device that allowed the tribes to condition their grazing strategies on its location, thereby avoiding conflict without direct communication. This focal point effectively transformed the game by correlating strategies and expectations, stabilizing a peaceful equilibrium.

$$
\begin{array}{|l|c|c|c|}
\hline
& G & NG & GiS, NGiN \\
\hline
G & 0,0 & 2,1 & 1,0.5\\
\hline
NG & 1,2 & 1,1 & 1,1.5 \\
\hline
GiN, NGiS & 0.5,1 & 1.5,1 & 1.5,1.5 \\
\hline
\end{array}
$$

Guala’s “Grazing game” (2015): strategies are Graze, Not Graze and Graze if North / Graze if South

This notion of observability crucial here was interpreted by Sugden and Zamarrón (2006) as focal points being underpinned not by instrumental but *pragmatic rationality* which is induced from its empirical success. Schelling stressed the “perceptual and suggestive element in the formation of mutually consistent expectations” (Schelling 1980, 83–84) which is essential for focal points, as the authors note. It means that there are no logical restrictions on converging for expectations between players. Herrmann and VanDrunen (2022) give a vivid example:

“In a signalling game, when a sender sends a signal to her partner, the partner responds to that signal in a specific way. But in the real world, her partner might not know what part of the act the sender performs is meant to be the signal. Suppose she signals by waving a red flag. What is the signal here? Is it the colour? The fact that it is a flag? The pattern in which she waves it? Where she stands when she is waving it?” (Herrmann and VanDrunen 2022, 2).

To make this problem easier, Sugden and Zamarrón suggest three features of Schelling’s focal point analysis:

* the players’ presumption of the solvability of coordination problems, which is based on empirical success of real agents
* players’ reasoning in finding focal points is not guided by principles of truth and validity essential for deductive systems
* players reason fully rationally in finding focal points.

As Schelling wrote:

“most situations - perhaps every situation for people who are practiced at this kind of game - provides some clue for coordinating behavior, some focal point for each person’s expectation of what the other expects him to expect to be expected to do” (Schelling 1980, 57).

The inherent difficulty with formalizing the idea of focal points consists of the “unstructured and unbounded set of possible coordination devices” (Sugden and Zamarrón 2006, 616), which Herrmann and VanDrunen (2022) illustrate with their red flag example. It means that any source of focality might present a theory in itself.

The notion of “cognitive selectivity” can bw helpful here. Mussweiler, Todd, and Crusius (2012) note that cognitive selectivity is an evolved characteristic of human cognitive system that plays a major role in individuals’ successful coping with inherently complex tasks like coordination. It is a fundamental mechanism of sense-making and navigating social worlds.

Selectivity is a feature of social information processing happening on three distinct stages:

* attention determines what information comes into the cognitive system,
* then this information is related to the already “stored” one
* and then a behavioral response is produced.

Stimuli requiring instantaneous behavioral response and contextually distinct stimuli tend to become more relevant and attended. And this is fair for both animals and humans.

While epistemic salience has to do with *beliefs*, natural salience has to do with attending to properties of environment and correlating strategies on them. Natural salience does not require any rationality at all, for agents update their behavioral strategies according to a simple rule like reinforcement learning. This presupposes few more deeper problems like how agents stabilize on a certain feature of environment as salient (Herrmann and VanDrunen 2022) and how they “decide” on who takes a sender or a receiver role (VanDrunen and Herrmann 2023)?

Although these types of salience and game-theoretic machinery behind them might seem unrelated and suited for studying different things, I will argue that they are deeply connected in the context of *evolution* of social conventions: how do we get epistemic salience from natural one. However, to arrive at a notion of *evolution* of social conventions and to be able to ask such questions, we need to frame the previous discussions of conventions within the context of social ontology and its naturalistic strand, in particular, which we will do in the next chapter.

## Interim results

In this chapter, we discussed the notion of *social convention* in the tradition from Hume to Lewis and modern game theory. We reviewed its intellectual precursors and influences, Lewis’s theory, its criticisms, extensions and larger-scope problem which his profound theory has generated:

* the relationship between norms and conventions and account of emergence of normative behavior Bicchieri (2005)
* the continuum between arbitrariness and functionality of conventions and the relationship between human and animal ‘conventions’ (O’Connor 2019, 2021; Harms 2004)
* the notion of common knowledge as a (not always) necessary condition for Lewisian conventions and the problem of agents’ cognitive architectures Peter Vanderschraaf and Sillari (2023)
* the problem of emergence of readily focal equilibria with epistemic and natural salience (Herrmann and VanDrunen 2022; Zachník 2021)
* the distinct equilibrium concepts of conventions: ESS and correlated (Skyrms 2014a; Peter Vanderschraaf 1995).

Crucially, these notions contain a subtle distinction between human and non-human conventions, illuminated from different angles. All in all, we can trace two models of rationality: low-demanding and high-demanding. Although there is no inherent problem with this, when game-theoretic thinking with its models is imported to other domains like social ontology, this subtle distinction between low-demanding and high-demanding rationality is often missing and generates controversies with conflation of two rationality models.

In the next chapter, we will put game-theoretic notion of convention into the context of *social ontology*, especially of a naturalistic kind, which is a strand of philosophy studying what there is in the social world. We will look at the standard notions proposed by J. Searle (1995), their refinements with game-theoretic conventions by Guala and Hindriks (2015) and arrive at a philosophical problem of evolution of social conventions underpinned by a deeper problem of *ontic reference* or *ontic account of scientific explanation* (Craver 2014). Afterwards, in chapter 3, we will sketch a pathway towards resolving the problem of ontic reference in game-theoretically inspired social ontology by building a model of evolution of human conventions from animal ones.

# **Chapter 2**. Problems of Guala’s unified social ontology

In this chapter, we shift gears to study the relationship between game-theoretic models of conventions, social ontology and how the former help *explain* or substantiate the latter in a viable social ontology.

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

A good deal of the current social ontology is detached from actual social science (Sarkia and Kaidesoja 2023). It has mostly been developed within analytic philosophy (Baker 2019; J. Searle 1995; B. Epstein 2015; Gilbert 1992) and social scientific tradition of critical realism (**archer1995?**; Elder-Vass 2010). There are views that social ontology is intrinsically more complex than any “natural” ontology for being “interactive”, or dependent on beliefs of individuals and hence changeable (Hacking 1999; B. Epstein 2015) and that social reality is inherently normative, meaning it cannot be descriptive and hence cannot import methodologies from “natural” ontologies (Ásta 2024; Haslanger 2012). Such a position renders and justifies the importance of an “extra-scientific” study of the social world. To me, these are two distinct projects: a set of *mutually coherent and explanatory powerful descriptions* of social phenomena and a normative theory thereof.

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

The difference between the strands is in the metaphysical relation of individuals and social facts. In the first case, social facts supervene on the facts about individuals, meaning that social properties cannot change without changing the individual ones. In the second case, facts about individuals *set up the conditions* for something to count as a social fact, i. e. dollar bills with a particular ink and paper and collective acceptance for money (2015, 58).

One of the most famous expressions of the latter is Searle’s (1995) formula “X counts as Y in context C”. For example, “bills issued by the Bureau of Engraving and Printing (X) count as dollars (Y) in the United States (C)“ (1995, 28). Epstein calls the former relationship (exemplified with”issued by”) *grounding* and the latter one (exemplified with “counts as”) *anchoring*, where grounding is responsible for *instantiation conditions* for being, for instance, money, in a particular context, and anchoring is responsible for *the mechanism of establishing instantiation conditions*. For example, in Searle’s case, *collective acceptance* of the fact X is such a mechanism[[20]](#footnote-64).

Such consideration exemplifies one of the two main views on the role of metaphysics[[21]](#footnote-65) in scientific inquiry. The first, empiricist one, renders metaphysics irrelevant and redundant or, if taken seriously, expects its constraints on models to impede the knowledge growth. The other view presupposes special metaphysical concepts beyond the vocabulary of science itself. For example, they discern grades of possibility and necessity largely unknown to science and go beyond causal dependence widely used in science (Ross 2023).

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

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

Social science and social ontology do not enjoy straightforfardness of mathematical formulations of its concepts like physics does[[22]](#footnote-66). Instead, it invites “folk” concepts like “beliefs”, “groups” and “social facts”. Further divide comes from the attitude towards this ambiguity. Some scholars argue that such “folk” ontologies are sufficient (**thomassen2003?**) and others put an interplay of “folk” and social scientific ontologies to the fore by bridging what (**sellars1962?**) called “manifest” and “scientific” image of phenomena within their social ontological theorizing (**asta2012?**; Guala 2016; **haslanger2018?**)[[23]](#footnote-67). The ambiguity between “folk” and scientific social ontology might indeed depend on mathematical apparatus and its underdeveloped character in social science. Ross (2023) also notes that ontology construction most probably will require its formulation with mathematical frameworks sensitive to structural modeling (like hierarchical Bayesian inference[[24]](#footnote-68)) and not set theory and similar tools from analytic metaphysics.

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

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

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

* reflexivity
* performativity
* collective intentionality.

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

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

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

Here, our classifications of the social world help establish and maintain it, whereas non-social objects are indifferent to our classifications of it, as Hacking (1999) puts forward with his distinction of interactive and indifferent kinds[[25]](#footnote-69). Nature’s objects do not change their behaviour given these classifications of them as opposed to social objects. This idea illustrates the notions of reflexivity and performativity characteristic for the “Standard model”. If social entities are comprised of beliefs about beliefs, their nature depends on these beliefs, and if beliefs change, social entities change accordingly. If social entities depend on beliefs about them, it is needed to constantly perform those to maintain them. To do this, individuals need to have collective intentionality about these beliefs. For example, for money to be itself, a relevant community has hold a collective intention to believe that certain physical entities can be used as a medium of exchange.

Along with the “Standard model”, Guala (2007) calls game theory one of the two main approaches to social ontology. He departs from the works of Schelling (1980) and Lewis ([1969] 2008), claiming that the latter was the first to apply rational choice analysis to social ontology and to present social conventions as systems of self-sustaining beliefs. Schelling, before Lewis, already was aware that coordination problems are unsolvable from the pure game-theoretic perspective if the existence of focal points is admitted, for the latter need a separate explanation. Schelling advocated the supplementation of an a priori approach to game theory with empirical studies of the emergence of focal points. This point was further explicated by Sugden and Zamarrón (2006), as we saw in the previous chapter.

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

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

Frank Hindriks and Guala (2015) propose a theory that social institutions like money, marriage or others can be seen as (sets of) conventions in Lewisian sense with added normaivity. They employ many arguments from our previous discussions like Aumann/Vanderschraaf *CE*, Skyrms/Maynard-Smith *ESS* and emphasize the role of *normative belief* in a manner similar to Bicchieri/Gintis. As we will pose a problem inside their theory later, it is important to carefully present and unpack it, which I will do in the next section.

## Guala’s and Hindriks’s “rules-in-equilibria” theory

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

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

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

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

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

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

Hindriks (2005) has challenged several aspects of Searle’s constitutive-rule framework. His deconstruction focuses on the notions of *status functions*, *Status Function Declarations*, and the role of language. Broadly, Hindriks argued that many of Searle’s theoretical apparatus are unnecessary or misleading, and that a more streamlined account can be given by focusing on collective acceptance and normative powers.

* **Status Functions as Deontic Powers.** Hindriks finds the term *status function* confusing and somewhat redundant. He suggests dropping the “function” and simply treating statuses as normative powers. In his words, “we can do without the term function while retaining the term status,” instead explicating statuses directly as the bundle of deontic powers they grant (**hindriks2009?**). By equating statuses with deontic powers (rights, obligations), Hindriks makes the normative dimension of institutions explicit, rather than hiding it under the metaphor of a “function” (**hindriks2009?**). Indeed, Searle himself has acknowledged that “all status functions are deontic powers,” which supports Hindriks’s move toward a more direct terminology (**hindriks2009?**).
* **Redundancy of Status Function Declarations.** Searle’s idea of a *Status Function Declaration* – a speech act that supposedly creates or recognizes a status – is, for Hindriks, unnecessary. He argues that the two key claims Searle attributes to such declarations (that collective acceptance is necessary and sufficient for the status) are already implicit in the standard “counts-as” formulation. Once we accept that *institutional statuses require collective acceptance, and that collective acceptance alone brings them into being* (the “Collective Acceptance Principle”), the special notion of a Status Function Declaration adds nothing new (Frank Hindriks and Guala 2015). Introducing declarations suggests without argument that only explicit speech acts can create institutions; Hindriks finds this unjustified and unhelpful. He concludes that Searle’s extra machinery (the Status Function Declaration with its “double direction of fit”) should be abandoned since it “does not add anything of value” (Frank Hindriks and Guala 2015).
* **Linguistic vs. Normative Distinction.** Hindriks also questions Searle’s emphasis on language as the source of all institutional power. In his earlier work, Hindriks has argued that the regulative/constitutive distinction is mainly a grammatical one: regulative rules are phrased with explicit imperatives or deontic terms, while constitutive rules are phrased with the “counts-as” locution, but both embed the same normative content (**hindriks2009?**). Normative obligations figure explicitly in regulative rules (“Do X” / “If Y do X”), whereas constitutive rules imply those obligations without stating them overtly. Thus, the locus of normativity is not really different between the two; only the linguistic presentation is. Hindriks calls for a view of institutions that centers on *collective commitment* and acceptance of standards, rather than on linguistic declarations per se (**hindriks2009?**).

In sum, Hindriks dismantles Searle’s superstructure of status functions and declarations, proposing instead that we should “explicate statuses in terms of normative powers” and rely on a simpler collective-acceptance principle (**hindriks2009?**; Frank Hindriks and Guala 2015). On this view, institutions are upheld by groups collectively endorsing certain rules, and the resulting normative powers of those rules are what really matters. His critique paves the way for unifying Searle’s approach with more analytical models, by translating constitutive claims into the language of regulative rules and equilibria.

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

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

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

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

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

Crucially, Guala reframes Searle’s project in game-theoretic terms without surrendering its insights. He acknowledges that institutional statuses influence how we classify and act, but maintains that these statuses are always aligned with behavioral regularities expressed as CE. Thus Searle’s emphasis on “counts-as” locution can be recovered as am epiphenomenon of coordination: an apparent creative *power of language*[[26]](#footnote-70), but in fact nothing more magical than the effect of correlated strategy profiles. As Guala & Hindriks put it:

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

A key task is to explain normative powers under this account. Guala & Hindriks argue that normative aspects can be modeled as payoff-modifications of the underlying game akin to -parameters of Crawford and Ostrom (1995). A normative rule adds incentives or penalties that make certain actions like cooperation more attractive. In practice, this means that if agents gain a “right” or incur an “obligation,” we represent this by inserting costs or benefits into the payoff matrix. This transforms a general-sum game into a coordination game where the efficient equilibrium becomes more salient.

Guala shows that adding such normative costs can create new equilibria that were not present before, or make the socially optimal outcome stable. Importantly, these modifications do not require a distinct ontological category beyond standard game-theoretic tools. Normative powers are simply part of the equilibrium framework: they enable coordination by altering incentives[[27]](#footnote-71). Guala & Hindriks illustrate that any status rule such as rights to use, transfer, or exclude can be recast as regulative rule once normative powers are included.

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

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

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

## Criticisms of Guala’s unified social ontology

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

* **Irreducibility of constitutive rules**. In a direct reply, J. R. Searle (2015) insists that constitutive rules cannot be eliminated. He warns that Hindriks & Guala’s project “only has a chance of success” if constitutive rules can be reduced to regulative ones and he flatly denies such a reduction. According to Searle, the core question of social ontology concerns status functions, which their account fails to capture. He argues that the equilibrium and regulative frameworks “cannot even pose” the right questions about institutions, and that only the constitutive-rule approach can supply answers. I am not convinced by this criticism, for Searle’s conceptual-analytic paradigm prevents him from *dynamic* analysis of social ontology, which implies questions of emergence of certain parts of his theory. Although I partially agree with Searle that regulative framework “cannot even pose” right questions, I see its cause in *different foci* of the two theories. Searle is concerned with performative *institutional facts* which do not necessarily solve a coordination problem. It works like this due to the very formulation of Searle’s position of epistemologically objective treatment of ontologically subjective reality. Social ontology is by definition dependent on representations and changes thereof by language. There is no room for “objective” ontological structures which Guala and Hindriks try to accommodate with conventions as beneficial behavioral patterns.
* **Neglect of material and historical aspects**. Rabinowicz praises Guala’s broad integration but raises concerns about the equilibrium focus. He points out that treating institutions as “rules-in-equilibrium” can misrepresent their ontology by ignoring material substrates and history (Rabinowicz 2018). Many institutions like universities, currency systems, markets and traditions involve *concrete people, practices and goods*, not just abstract strategies. Material base of the games played like students and classrooms in a university, as Rabinowicz claims, is essential part of institutions and cannot be abstracted from due to disregarding “basic ontology” of the modeled phenomena. This criticism fails to convince me, as I find the notion of the “basic ontology” misleading due to implying some “folk ontology” and subscription to conceptual analysis which favors intuition of the social world as a valid starting point for ontological scrutiny. Models abstract away inessential parts of systems in question. Although I agree that there are always concrete people and practices with their material bases, including misleading “basic ontology” is inessential for explanation of the structure of social ontology, emergence and stability of its core parts. Another Rabinowicz’s criticism is that the set of possible CE is larger than that of NE, thus complicating the problem of equilibrium selection instead of making it easier. IT is so for there is infinity of possible probability distributions which can create many CE arrangements. However, seen *dynamically*, as Skyrms (2014a) does, for example, correlation of strategies arises in the course of evolution, thus solving the problem of equilibrium selection with symmetry-breaking by chance and stochasticity.
* **Irreducibility of normativity**. (**roversi2021?**) defends the traditional importance of constitutive rules. He finds the unified account too reductionist: treating constitutive rules as summaries of regulative incentives “strips away” their normative essence (**roversi2021?**). From his perspective, this view fails to explain why individuals feel bound by institutional norms beyond calculative incentives. Roversi insists that one cannot capture the meaning of a status (like professor) by conditional rules alone as they imply a more complex notion of normativity. Contra Roversi, I believe that Guala and Hindriks to not want to *reduce* normativity to instrumental rationality. On the opposite F. Hindriks (2019) elaborates on the definition of social institutions as norm-governed social practices and contends that modeling social norms as sanctions with costs that agents incur for violating norms, is insufficient for its perception by agents as legitimate, According to Hindriks, instrumental account fails to capture the motivation by the norm itself and not by the costs of its violation. He claims that it is *normative expectations* and especially *normative beliefs* that complement sanctions as a source for norm existence and perception as legitimate. Social norm governs a practice if its participants are motivated to follow its rule to a noteworthy extent. This view partially departs from RiE and Hindriks even refines RiE as “rules *and* equilibria”, thus disentangling normativity from equilibria.
* **Overestimation of coordination**. Hedoin (2016, 2021) has criticized RiE on the grounds of misrepresentation of institutional stability as overly based on coordination. While Guala’s game-theoretic theory offers an elegant and parsimonious model of institutions as coordination equilibria, it is ultimately too thin to capture the full social ontology of institutions. Hédoïn’s constructive thesis calls for a broader, normatively rich, and historically grounded framework that recognizes institutions as constitutive of agents’ preferences and identities, embedded in social and cultural contexts, and endowed with normative powers that generate genuine obligations. This richer approach aims to explain not just how institutions solve coordination problems but how they shape social life at a fundamental level. The main novel idea is to make exogenous preference endogenous, meaning that equilibria do not reflect agents preferences, but *constitute* them. For example, the institution of marriage changes how individuals value relationships, not just how they coordinate behavior. This dimension is missing in Guala’s payoff-matrix approach. While this is true, Hedoin’s argument weakens in the light of the dynamic view of institutions: both coordination problems and values might have evolved as adaptations to each other, and as Ross (2012) noted, it is hard to say what was first. It means that Guala’s theory can accommodate the transformative view of institutions by integrating values into game matrices and saying that institutions change the rules of the game as Vlerick (2016) does. In addition, Frank Hindriks and Guala (2021) argue that one of the main functions of institutions is to preserve values by the means of social norms, which means that RiE at least presupposes integration of values.

Although the aforementioned critiques highlight important drawback of RiE, nost of them do not go deep enough to spot a problem in the very structure of RiE which affects its overall explanatory capacity. As I will argue in the next section, there are intrinsic problems with the core notions of “correlation” and “representation” within RiE which make the introduction of rules and entire explanation based on the comparison of game-theoretic models in human and animal cases problematic.

## Problems with “correlation” and “representation” in RiE

What is the justification for connecting rules and equilibria accounts of social institutions? According to (**guala2016b?**), institutions as rules-in-equilibria are normatively-driven behavioral regularities comprising CE. “Rules” here are the recipes guiding and prescribing certain behavior and are used by the agents themselves, and ”equilibria” are *objective stable states* of the strategic interaction between agents. Other scholars pinpoint normative and self-sustaining nature of institutions as “humanly devised constraints that shape human interactions” (North 1990), “norm-governed social practices” (Tuomela 2013) and “self-sustaining salient behavioral patterns” (Aoki 2007). Institutions combine “subjective” and “objective” components: they are driven by social norms, that might vary from one population to another, and, at the same time, constrain possible actions and sustain itself.

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

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

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

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

$$
\begin{array}{|c|c|c|}
\hline
& C & D \\
\hline
C & -1,-1 & -3,0 \\
\hline
D & 0,-3 & -2,-2 \\
\hline
\end{array}
\quad
\begin{array}{|c|c|c|}
\hline
& S & H \\
\hline
S & 4,4 & 1,3 \\
\hline
H & 3,1 & 2,2 \\
\hline
\end{array}
$$

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

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

$$
\begin{array}{|c|c|c|c|}
\hline
& H & D & B\\
\hline
H &-1 &2 &0.5\\
\hline
D &0 &1 &0.5 \\
\hline
B &-0.5 &1.5 &1.0 \\
\hline
\end{array}
$$

“Hawk-Dove-Bourgeois” game

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

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

$$
\begin{array}{|l|c|c|c|}
\hline
& G & NG & GiS, NGiN \\
\hline
G & 0,0 & 2,1 & 1,0.5\\
\hline
NG & 1,2 & 1,1 & 1,1.5 \\
\hline
GiN, NGiS & 0.5,1 & 1.5,1 & 1.5,1.5 \\
\hline
\end{array}
$$

Guala’s “Grazing game” (2015)

The payoff structure of the game is identical for animal and human cases, and we cannot discriminate between them solely on this basis. This invites the insufficiency of CE employed by Guala and Hindriks. However, although they mention adding payoff modifiers to the game structure to accommodate inherent normativity of conventions (F. Hindriks 2019), they do not explain how animal conventions evolve into human payoff-modified ones[[28]](#footnote-74). So the question is twofold:

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

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

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

To summarize Guala’s and Hindriks’s argument:

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

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

### The “scope of actionable signals” in more detail

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

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

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

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

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

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

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

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

If we abstract away from a particular case of riverbed-based borderlines, the amount of things which could serve as a correlation device for coordinating on borders and property would grow, for there is *functional independence* of human correlation devices from a set of environmental features which could possibly be salient. We can *use* different things as correlation device in functionally similar situations like deciding who goes first when pedestrians or cars clash into each other. There is a possibility, however, that there is an exhaustive set of functionally possible correlation devices which can be deployed across contexts varying from more mechanistic like “precedence” and “spatial position” to more social constructed categories like “age” (since its perception and attitude towards can vary culture-wise), “gender”, “role” and others[[29]](#footnote-75).

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

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

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

Thus, *environmental asymmetries appearing to organisms as salient comprise a set of possible correlation devices*. The greater the number of recognizable asymmetries, be it physical or epistemic, the greater the amount of possible correlation devices. So, what Guala most probably means by “scope of actionable signals” is essentially a set of recognizable salient environmental asymmetries. We will use this insight later in Chapter 3 when modeling the cognitive evolution of correlation devices.

### Conflation of agency models in RiE

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

* coordination
* correlation
* representation,

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

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

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

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

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

And ESS is given by:

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

* either “animal conventions” are mathematically different from human social institutions, for they represent ESS and not CE, and hence Guala’s argument from insufficiency and the very reason for introduction of rules seems to fail (for one misinterprets ESS for CE with its rationality assumptions);
* or “animal conventions” are themselves correlated, and there comes the need of showing *in what respect these correlations are different*. Only by showing it we can disambiguate the claims of RiE about CE as its backbone.

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

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

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

### Ontic and epistemic correlation of strategies

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

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

$$
\begin{array}{|l|c|c|}
\hline
& \textbf{Hawk} & \textbf{Dove} \\ \hline
\textbf{Hawk} & $\frac{(V-C)}{2}, \frac{(V-C)}{2}$ & $V, 0$ \\ \hline
\textbf{Dove} & $0, V$ & $\frac{V}{2}, \frac{V}{2}$ \\ \hline
\end{array}
$$

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

This model assumes:

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

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

* either or
* and .

According to these conditions, cannot be an ESS, for , and is an ESS if costs of injury are less than potential gain from the resource, . If , neither nor 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 with probability , and with probability ’.

A mixed strategy , 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 and Cannings (1978): if a mixed ESS includes the pure strategies with non-zero probability, then

Intuitively, this means that if , adopting more often and less often would be more advantageous than following strategy , making it not an ESS.

However, can be an ESS if probability of its adoption is . In contests where the cost of injury is greater than the rewards of victory, , mixed strategies with 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.

$$
\begin{array}{|l|c|c|c|}
\hline
& \textbf{Hawk} & \textbf{Dove} & \textbf{Bourgeois} \\
\hline
\textbf{Hawk} & 0,0 & 2,1 & 1,0.5 \\
\hline
\textbf{Dove} & 1,2 & 1,1 & 1,1.5 \\
\hline
\textbf{Bourgeois} & 0.5,1 & 1.5,1 & 1.5,1.5 \\
\hline
\end{array}
$$

Hawk-Dove-Bourgeois game payoff matrix

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

Here, the player chooses and with equal frequency, acting as an owner on half the occasions and an intruder on the other half. And when two ’s compete, if one chooses , the other chooses . If , the ESS is as it is worth risking injury to gain the resource; if , the ESS is as ownership settles the contest without escalation. It means that in both ‘Hawk-Dove’ and ‘Grazing’ games .[[30]](#footnote-78).

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

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

It is interesting that Maynard Smith (1982) considers the ‘social contract’ game as one which humans can play but animals cannot. This game involves a group of individuals agreeing on a behavioral regularity and punishing any member who deviates from it, thus implying “full-fledged”a normativity. 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.

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

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

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

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

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

One intrinsic problem with as CE, however, is the source of randomization. Some scholars appeal to Nature as to such source and correlation device in the animal case, thus eliminating the tension between the requirement of randomization and symmetry of ESS (Cripps 1991; Skyrms 2014a; Metzger 2018). In particular, Gintis defines CE as an NE of a game augmented by the *initial move by Nature* that who observes a random variable on a probability space and issues directives to each player , such that choosing the directive is the best response given agents having a common prior and assuming other players are also following Nature’s directives (Gintis 2009b, 135–36). The crucial assumption, though, is that *the game is epistemic and has common priors*. This implies that all agents agree on the probability distributions over their actions and have *joint randomized probability*. Assuming a strict CE—that each agent has a single best response in every state —each player’s move is known to the others and rationality dictates that each must play a best response to the actions of the others. If we plug here our earlier example of sets of correlation devices, we can represent as a composite cue vector which encode all the relevant information about the world state.

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

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

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

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

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

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

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

We might introduce the distinction of *ontic* and *epistemic* correlation. What is subtle yet crucial here and what is overlooked in RiE, is that these types of correlation have distinct real-world references. While both animal and human conventions are CE formally and can be described with a single model, both their “inner workings” and references “in the world” are different. This renders the entire Guala’s argument of “scope of actionable signals” problematic, for it is not evident how to compare games with ontic and epistemic cues. To remind, this argument is at the backbone of RiE as it justifies the introduction of rules into equilibrium theories of social institutions.

According to Guala and Hindriks, stability of social institutions cannot be based only on CE as it is equally applicable to animals and humans. However, these are two distinct CEs — frequentest (ontic) and Bayesian (epistemic) ones. While rules or any notion of normativity is missing in the ontic correlation by definition (as it cannot be a feature of environment[[31]](#footnote-79)), rules as salient norms can and sometimes are CE themselves. For example, Gintis (2009b) views social norms as “choreographers” and CE as they are essentially “second-order” conditional strategies. We can imagine a pair of strategies like “Others expect me to stop if red” and “Others expect me to drive if green” which comprise a CE.

It means that Guala’s and Hindriks’s argument for insufficiency of CE without rules for human cases is unsubstantiated, for CE as norms already includes both Bayesian rationality *and* normativity (for example, Gintis’s belief matrices from the previous chapter). Consequently, explanation of stability of institutions might nit require separate rules as norms, as for epistemic games with conditional agents, norms can be CE themselves.

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

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

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

flowchart LR  
1(Representation) --grounds--> 2(Normativity) --grounds--> 3(Social stability)

Conceptual relations of main explanatory components of RiE

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

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

### The problem with “representation” in RiE

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

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

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

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

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

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

There are two types of rules in RiE:

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

Strategies of the game can be described as rules, but a-rules influence behavior and o-rules only describe it (**guala2016b?**).

*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.

flowchart LR  
 subgraph Objective  
 3(Observer rules)  
 1(Correlated equilibrium)  
 1<-.->3  
 end  
 subgraph Subjective  
 3<-.Representation.->4  
 end  
 4(Agent rules)

Institutional components in rules-in-equilibrium theory

What Guala and Hidriks propose with introduction of representation as a necessary component of a theory of social institutions is essentially a way of connecting objective equilibria and subjective rules. And a core idea is that the latter can misrepresent the former as I have already noted: agents can have false beliefs about objective states.

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

The constitutive rules approach argues that our beliefs are essential for the existence of institutions, which involve more than just actions. This applies to objects, persons, and events too. X in “X counts as Y in C” can be replaced by predicates that refer to any ontological category (Guala and Hindriks 2015, 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 and Hindriks 2015, 477).

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

From a scientific point of view, representation as a relation makes sense, for it allows investigation of Y-terms like “money” used by agents (**guala2016b?**, ch. 14). However, representation as a cognitive capacity does not have any immediate practical application, especially in sociological data. Hence, we can discern two notions of representation in Guala’s and Hindriks’ argument:

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

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

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

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

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

The argument presupposes that behavior conditioning, and hence strategy selection, occurs already being based on existing rules. To follow a rule, it should already exist in the structure of the game. At the same time, these rules are a-rules, and they already represent existing strategies “that ought to be followed in a given game”. It means that behavior is conditioned on existing strategies, and this involves something like a reductive regress: inventing new rules requires not only a capacity for stimulus-behavior decoupling, but existing equilbria, for here salient features of existing equilibria are used as coordination devices. Although it can be problematic only from a logical point of view and not from a pragmatic and evolutionary one, since realistically we cannot think of a situation with no strategic equilibria (maybe before life?).

In other words, the authors equate representation of salient features of the environment with representation of existing strategies, or behavioral responses, that 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, which does not allow them to form whimsical private (mis)representations.

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

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

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

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

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

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

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

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

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

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

## Representational explanation as source of problems in RiE

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

According to a prominent family of views in the philosophy of science, scientific explanations fundamentally involve subsuming a phenomenon or its description under a general representation. This *representational subsumption view* (RSV) has been articulated in various forms: for (**hempel1965?**), explanations take the form of generalizations expressed in first-order logic; Kitcher (1981, n.d.) conceives them as argument schemas; Bechtel and Abrahamsen (2005) emphasize mental models; Churchland (1992) appeals to prototype vectors[[32]](#footnote-82); and Machamer, Darden, and Craver (2000) focus on mechanism schemas. Despite differences in the specific representational format, these accounts share the core assumption that the philosophical debate about scientific explanation should center on the nature of these representational forms.

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

The distinction between ontic and representational accounts of scientific explanation shows a fundamental philosophical divide in understanding how scientific explanations function as they have diverging underlying assumptions. The ontic perspective posits that explanations are not primarily linguistic or cognitive entities but exist in the objective causal structures of the world itself. From this viewpoint, a philosophical theory of scientific explanation must achieve two critical goals: explanatory demarcation and explanatory normativity. The former distinguishes explanation from other scientific achievements like “control, description, measurement, prediction, and taxonomy,” while the latter determines “when putative explanations succeed and fail” (Craver 2014). The proper question to models on ontic account should be “what our models should represent?”.

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

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

## Interim results

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

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

We spotted a problem with agency models in HDB game where Guala and Hindriks illegitimately transferred CE from evolutionary context to an epistemic one. To reiterate, animal HDB in Maynard Smith from whom Guala and Hindriks take the notion of HDB, although it can be seen as CE, *operates on phenotypes* or non-epistemic agents and not boundedly rational Bayesian individuals as Guala and Hindriks propose. This conflation generates a problem in their insufficiency argument for introduction of rules as separate explanatory entities: in human HDB, normativity is already there, while in animal HDB rules are just not needed as a correlation device. I argued that this generates a further problem with introduction of the notion of “representation” in RiE as an explanation of the difference between animal and human conventions, for Guala and Hindriks seem to conflate scientific and cognitive notions of representation which leaves normativity unexplained and the gap between animal and human conventions mysterious. Importantly, I suggested that the “scope of actionable signals” can be quantified as a set of potentially salient environmental features which can be used as correlation devices, and that humans have more functionally abstract correlation devices. I also argued that problems of RiE might be due to the authors’ (probably implicit) subscription to a representational subsumption view of scientific explanation, according to which the explanatory power of a model comes from its representational features rather than from its real-world counterpart of entities and causal processes responsible for an outcome.

What we do next is building models to (try to) answer the question of how cognitively sophisticated and functionally abstracted correlated devices might have evolved from non-normative and non-epistemic agents playing CE in HDB. We are addressing two weaker yet explanatorily crucial parts of RiE: relationship between animal and human conventions as CE and emergence of normative human conventions with decoupled representation.

# **Chapter 3**. Evolution of normativity in Guala’s unified social ontology from non-epistemic agents with niche construction

As we saw in the previous chapter, Guala and Hindriks explain the stability of social institutions with both objective and subjective parts: equilibria of strategic games and meaningful rules employed by agents with which they represent the objective equilibria of games. I claimed that Guala and Hindriks bet too much on an ambiguous notion of representation as responsible for normativity which, according to them, stabilizes institutions. The ambiguity stems from the description of representation as correspondence between scientific concepts and its real-world referents, while using cognitive capacity language. Discerning these notions shows that scientific notion of representation is what makes RiE work, but scientific representation depends, as I claimed, on *decoupled representation* as evolved cognitive capacity to disentangle stimuli and response and internalize norms. Showing how a representational capacity contributes to rule-following is central for our argument.

A reasonable place to start is to unpack Sterelny’s theory of decoupled representation mentioned by Guala and Hindriks in their appeal to cognitive representation as responsible for coordinating action on many representations as following rules. Taking Sterelny’s theory seriously will allow us to see that its core mechanism of *cognitive niche construction* is ecological, meaning that the novelty is introduced through the interaction of agents with their environment. This will provide an important clue for resolving both problems of RiE: emergence of epistemic correlation from ontic one and emergence of normativity. A capacity of agents to modify the informational structure of their environment affects the ways of responding to environment, thus providing a “ratcheting” mechanism of what Sterelny calls “downstream niche construction”.

Next, we look into Sterelny’s theory of decoupled representation, its main mechanism of cognitive niche construction, game-theoretic approaches incorporating niche construction and arrive at our own agent-based model of emergence of epistemic correlation and normativity.

## Sterelny’s theory of decoupled representation

Guala and Hindriks mentioned Sterelny’s notion of decoupled representation as a key differentiator of agents capable of larger “scope of actionable signals” ir what we called arbitrary and functionally abstracted correlation devices. We spotted that this mention is explanatorily insufficient for substantiating normativity and “rules” as decoupled representation is only responsible for flexible and context-relevant usage of “stored” representations of external stimuli detached from direct response. The claim that conditioning behavior on many representations is the same as following rules is not right, for normativity does not come from many representations alone. Most probably it comes from the different *nature* of them, depending on the level of their *functional abstraction*.

Sterelny has offered a comprehensive evolutionary framework for understanding the origins and nature of human cognition (Sterelny 2003, 2012b). Central to his account is the concept of *decoupled representation*, which refers to internal cognitive states that represent aspects of the environment independently of immediate behavioral responses. This concept is pivotal in explaining how humans evolved the capacity for flexible planning, social cooperation, and cumulative culture.

Sterelny situates cognitive evolution within a “hostile world” - an environment marked by unpredictability, competition, and complexity. In such a world, simple stimulus-response mechanisms are insufficient; organisms benefit from representations that are *decoupled* from immediate stimuli and rigid behavioral patterns, enabling flexible and strategic behavior. In a way, decoupled representation is responsible for emergence of *individual strategic behavior* describable by game theory as Lewis, Guala, Vanderschraaf, Gintis and others have been doing. This fact highlights the distinction of models with phenotypes and Bayesian individuals as agents within RiE which I noted in the previous chapter as the former do not possess capacities for strategic behavior and the latter do.

It is important to understand the conceptual foundations of Sterelny’s theory to fully apply it to RiE. His theory is deeply interdisciplinary, drawing on evolutionary biology, philosophy of mind, and anthropology. It builds on and refines earlier philosophical work, particularly Millikan’s biosemantics (1987) and Godfrey-Smith’s theory of cognition as functional adaptation to environmental complexity (Godfrey-Smith, n.d.). Moreover, Sterelny integrates niche construction theory (Odling-Smee, Lala, and Feldman 2003), emphasizing how organisms modify their environments informationally in ways that feed back on cognitive evolution.

Probably, the most foundational to Sterelny’s theory is Godfrey-Smith’s Environmental Complexity Thesis (ECT), a quite important idea in the philosophy of biology and cognitive science. The core claim of the ECT is succinctly stated:

“The function of cognition is to enable the agent to deal with environmental complexity” (Godfrey-Smith, n.d., 3)

Godfrey-Smith’s ECT asserts that the complexity of an organism’s cognitive architecture is shaped by the complexity of its environment. Specifically, cognition evolves as an adaptation to environments that are variable, unpredictable, and structured in ways that require flexible, context-sensitive responses. This is a form of what Godfrey-Smith calls “c-externalism” (“c” means complexity): the idea that internal complexity in organisms can be explained in terms of the complexity of their external environments (Godfrey-Smith, n.d., 4). There are several core claims connected to ECT:

* *Necessity, not sufficiency*: the ECT claims that environmental complexity is necessary for the evolution of cognition, but not sufficient as other factors may play a role. It is similar to how decoupled representation is necessary and not sufficient for normative rule-following in RiE
* *Function and adaptation*: cognition is understood *teleonomically*. Its function is the effect that explains its presence, typically as a result of natural selection favoring agents that can mediate well-adapted responses to environmental variability. This is very much like Wright’s notion of biological function (Wright 1973) employed by Guala and Hindriks in their justification of functions of institutions (Frank Hindriks and Guala 2021)
* *Behavioral and mechanistic levels*: the ECT bridges behavioral and mechanistic explanations, relating the functional properties of behavior to environmental structure, and offering a framework for understanding how agent-environment systems operate.

For Godfrey-Smith, both abiotic and biotic factors can generate environmental complexity: climate, terrain, other organisms and especially conspecifics. The latter is particularly important in social species, where the *activities of others create a “causal cycle” of complexity and response*: organisms both shape and are shaped by their social environments (Godfrey-Smith, Sternberg, and Kaufman, n.d.). A similar idea concerning specifically human evolution is due to Zawidzki (2013) and his idea of “mindshaping” and complementing mindreading, which area cognitive capacities to recursively reason about other’s reasoning akin to Lewis’s “common knowledge” or “common reason to believe”. This recursive dynamic is crucial for understanding the evolution of cognition in social animals including humans. Note the “causal cycle” as it will be of great import for our model.

Godfrey-Smith also carefully distinguishes between “constructivist” and “externalist” explanations. While externalist explanations account for organismal properties in terms of environmental features, constructivist explanations focus on how organisms actively modify their environments. Godfrey-Smith’s “narrow constructivism” requires that *organisms causally intervene in the external world*, not merely change themselves in response, as Bedau (1996) notes. This distinction is important for understanding the feedback loops between organisms and environments in cognitive evolution.

The ECT has been influential in both empirical and theoretical research, providing a framework for *agent-based modeling* and studies of behavioral flexibility (Seth 2002). It underpins Sterelny’s argument that increasing environmental and social complexity selects for more flexible, decoupled forms of representation, as organisms must cope with uncertainty, hazard, and the evolved strategic behavior of others.

Another foundational source of Sterelny’s decoupled representation theory is Millikan’s theory of signals, especially her distinction between *indicative* and *directive* (or imperative) signals (Ruth Garrett Millikan 1987; Ruth Garrett Millikan and Millikan 2004).

Millikan famously has distinguished between two classes of animal signals:

* *Indicative signals*: these are public representations whose function is to indicate or “tell the truth” about some state of the world. They are belief-like, corresponding to what is the case. For example, a vervet monkey’s “snake alarm call” indicates the presence of a snake
* *Directive (Imperative) signals*: these are signals whose function is to direct or motivate behavior. They are desire-like, corresponding to what should be done. For example, the same snake alarm call also functions to direct conspecifics to climb trees which is a specific “hard-wired” anti-predator response.

Millikan’s insight is that many natural signals and representations are “dual-purpose” having both indicative (informational) and directive (motivational) functions, but these *can be separated and evolve independently*, which can help explain the emergence of normativity “naturally”.

Millikan’s framework explains how signals and representations can evolve to serve different functions in animal communication and cognition. In simple organisms, signals are often tightly coupled: the indicative and directive functions are fused. For example, a chemical cue triggers a fixed action, like in ants. In more complex organisms, these functions can become disentangled, allowing for flexible, context-sensitive responses which is a key step toward the evolution of belief-desire psychology and decoupled representation.

Millikan’s theory dovetails with Godfrey-Smith’s ECT: as environments become more complex and unpredictable, there is selective pressure for organisms to evolve cognitive architectures that can separate the indicative (what is the case) from the directive (what to do about it) signals. This separation enables organisms to hold beliefs about the world independently of immediate action, supporting planning, learning, and, most interestingly for us, social coordination.

Together, these frameworks explain why and how cognition evolves to meet the demands of complex environments, and how the separation of indicative and directive functions underlies the emergence of flexible, decoupled representation at the heart of human cognition.

Sterelny’s analysis begins with identification of three ascending grades of representation:

1. **Detection Systems**:

* Found in simple organisms, these are tightly coupled stimulus-response mechanisms. For example, ants detect oleic acid as a single cue to remove corpses, triggering a fixed behavior. Here, representation and action are inseparable; the “indicative” (detection) and “imperative” (action) functions are fused;
* Such systems are adaptive in stable environments where cues reliably predict outcomes. However, in more variable or ambiguous environments, these rigid couplings become maladaptive, leading to inappropriate or inefficient behaviors;
* For example, a frog’s tongue-flick response to small moving objects is effective in simple settings but fails in cluttered environments where non-prey items move. Here, the frog’s detection system cannot discriminate context, leading to wasted effort.

1. **Robust Tracking**:

* Intermediate organisms integrate multiple cues to track environmental states more reliably. A bird might use visual, auditory, and olfactory signals to identify predators, reducing reliance on single, error-prone cues. This “input-side” flexibility enhances accuracy but does not yet decouple representation from specific outputs and make it flexible for offline use and interpretation;
* Sterelny emphasizes that robust tracking enhances accuracy by *pooling information* but does not fully decouple representation from behavior. The *behavioral repertoire remains limited* and relatively fixed. For example, the bird’s detection of a predator still triggers a narrow set of responses such as alarm calls or flight.

1. **Decoupled Representation**:

* The highest grade, seen in humans and some social mammals, involves internal states that track environmental features like predator presence without dictating fixed responses. These representations are *multipurpose*: the same belief about a predator’s location could inform fleeing, hiding, or attacking, depending on context. They provide a flexible informational basis for selecting among a range of possible actions, depending on context, goals, and internal states;
* This flexibility enables organisms to engage in *means-end reasoning*, planning, and strategic social behavior. For example, a chimpanzee tracking the dominance rank of a rival may decide to challenge, avoid, or ally with that rival based on its own strength and social context;
* Decoupled representations require significant cognitive resources, including *working memory* and the ability to maintain and manipulate internal states independently of immediate stimuli. Sterelny argues that such representations are metabolically costly and developmentally demanding, which explains their rarity outside humans and a few social mammals.

The *hostile world* in which Sterelny embeds the evolution of decoupled representation is an environment marked by unpredictability, competition, and complexity. Sterelny argues that such environments impose selective pressures favoring cognitive flexibility and strategic behavior. Key drivers include:

* environmental translucency
* social complexity
* niche construction.

The unpredictability, or *environmental translucency*, means that cues are often ambiguous or misleading. For example, a rustling bush might signify prey, predator, or merely wind. In such cases, organisms benefit from maintaining *internal representations that can be updated* and used flexibly rather than triggering fixed responses. We can already see parallels with Bayesian agents as having advantage over non-Bayesian ones, since updating internal beliefs and maintaining memory are beneficial for *information integration* which is crucial for adaptive learning.

Sterelny aligns with the *social intelligence hypothesis*, which posits that the cognitive demands of living in complex social groups drove the evolution of advanced cognition. He argues that decoupled representations initially evolved to track *social information* like alliances, hierarchies and intentions effectively enabling strategic social behavior.

In primate societies, individuals must monitor the shifting alliances and rivalries among group members. Decoupled representations allow an individual to maintain a *mental map* of social relationships that can be used flexibly to guide behavior, such as choosing when to cooperate, compete, or deceive.

Lastly, Sterelny incorporates *niche construction theory* (Odling-Smee, Lala, and Feldman 2003) to explain how organisms modify their environments in ways that feedback on cognitive evolution. He argues that early hominins constructed *cognitive niches*: socially and materially modified environments that *scaffolded* the development of decoupled representations. As he puts it:

“Niche construction creates new selective pressures and cognitive challenges. By modifying their environments, organisms create opportunities for planning, innovation, and cultural transmission that require flexible, decoupled cognition.” (Sterelny 2003, 89).

Examples include tool use, landscape modification, and *social institutions*, which extend cognitive capacities beyond biological limits. For instance, the use of fire or construction of shelters alters ecological conditions, creating new problems and opportunities that require planning, flexible behavior and, importantly, social coordination (Sterelny 2021).

These constructed niches provide stable contexts in which learning and cultural transmission can occur, further reinforcing the selective advantage of decoupled representation.

A crucial feature enabling the evolution of decoupled representation is *developmental plasticity*, especially as it interacts with processes of niche construction. Developmental plasticity refers to an organism’s ability to modify its behavior, physiology, or morphology in response to environmental conditions during its development. This plasticity is crucial for the evolution of cognitive architectures that support decoupled representation for several reasons:

* *Facilitating behavioral flexibility:* plasticity allows individuals to adjust their responses to a wide range of environmental stimuli, rather than being limited to fixed, hard-wired reactions. This flexibility is a necessary precursor for the emergence of representations that can be “decoupled” from any single cue or response, enabling organisms to use information in a context-sensitive, rather than stimulus-bound, manner.
* *Supporting general-purpose information storage:* decoupled representations serve as a general-purpose information store, allowing organisms to accumulate and deploy knowledge across different contexts. Developmental plasticity provides the substrate for this by enabling the learning and retention of information that is not immediately linked to a specific behavioral output, but can be flexibly applied as circumstances demand.
* *Enabling conditional and hypothetical reasoning:* with plasticity, *organisms can entertain conditional beliefs* “if X do Y, then Z will result” and use internal models to guide novel or indirect forms of action. This is a hallmark of decoupled representation, moving beyond direct stimulus-response links to support practical reasoning and planning. It is also, probably, the most important cognitive feature which Guala and Hindriks wanted to integrate into RiE. It is also worth noting that the distinction of ontic and epistemic correlation from Chapter 2 can be seen as distinction between ecological (or relational) condition like “if owner, play Hawk” and conditional *reasoning*. And the shift is from “environmental” probabilistic to individual Bayesian inference.

Developmental plasticity is especially potent when coupled with *niche construction*. Sterelny argues that niche construction and plasticity are mutually reinforcing:

* *Stabilizing and scaffolding cognitive development:* organisms can shape their learning environments through parental care, social learning, environmental modification or in some other way, making it more likely that plastic developmental processes will lead to adaptive, decoupled representations. In turn, plasticity allows individuals to take advantage of these constructed niches, further enhancing the reliability and utility of their internal models.
* *Driving the evolution of cognitive complexity:* as organisms construct more complex and information-rich niches, selection favors greater developmental plasticity and more sophisticated representational capacities. This co-evolutionary dynamic helps explain the emergence of belief-like, decoupled representations in lineages with extensive niche construction, such as humans.

As I have mentioned, plasticity supports conditional reasoning. It does so in several ways:

* *Adaptive neural investment:* variable environments select for neural systems that can flexibly encode and update associations between cues and outcomes. This means that, rather than being locked into rigid stimulus-response patterns, individuals can learn to map specific environmental signals (X) to context-appropriate actions (Y), and revise these mappings as circumstances change. Such flexibility is a definitive sign of conditional reasoning. An important difference between “X” here and in RiE is that in the latter it can be an abstract entity like a social norm, while in evolutionary scenarios “X” is more concrete and often physical feature of environment.
* *Learning from social and environmental cues:* plasticity allows individuals to use early-life experiences, social learning, and feedback to shape their cognitive strategies. For example, exposure to reliable cues of social support or adversity can calibrate neural development and bias individuals toward certain forms of decision-making or problem-solving. This developmental tuning enables agents to form conditional strategies such as “if signal A is present, cooperate; if signal B, defect” that are sensitive to the specifics of their environment.
* *Evolved probabilistic cognitive mechanisms:* plasticity underlies what evolutionary developmental psychology calls “evolved probabilistic cognitive mechanisms” which are information-processing systems that are not hard-wired but are expressed flexibly, depending on developmental context. These mechanisms allow for the emergence of domain-general reasoning abilities, including the capacity to counterfactual reasoning and updating strategies based on experience. This aligns with Guala’s view on conditional agents as boundedly rational.

In evolutionary models, agents may begin as simple phenotypes with fixed behavioral strategies. *Plasticity* enables some individuals to move beyond these fixed patterns, learning to condition their actions on salient cues and to adjust these condition-action rules as they gain experience. This transition is facilitated by developmental processes that allow for the formation and refinement of internal representations and expectations about the environment and other agents.

Plasticity also supports the uptake and internalization of social norms. Through imitation, tutoring, and feedback, agents can acquire conditional rules that are prevalent in their group (“if the light is red, stop; if green, go”), and these rules become stabilized as shared expectations. Over time, such conditional strategies are not just learned but become institutionalized as norms, further reinforcing their spread and stability.

Guala’s agents do not require perfect rationality or full-fledged Bayesian updating based on probability distributions. Instead, his agents use *bounded, heuristic updating*: agents adjust their conditional rules based on the observed success or failure of past actions, gradually converging on strategies that yield higher payoffs in their ecological and social niche. This presupposes a rather peculiar cognitive architecture probably joining *reinforcement learning* with elements of *Bayesian updating*, which we will consider later in this chapter.

Sterelny has developed his theory further by focusing on *apprenticeship learning* as a key mechanism enabled by decoupled representation. It is essentially a higher-level learning mechanism *in situ* which allows for structured social learning, where novices acquire skills through observation, imitation, and explicit instruction. Sterelny writes:

“Apprenticeship learning requires metarepresentational capacities—understanding the intentions and goals of teachers—and the ability to decouple representations of skills from immediate utility.” (Sterelny 2012b, 112).

Unlike many non-human animals, humans engage in *detached pedagogy*, where teaching is intentional and involves explicit communication about how and why to perform tasks. This requires the ability to represent others’ mental states (also called “theory of mind”) and to hold representations that are not immediately tied to current goals or stimuli, but are *abstracted* from immediate utility.

For example, a master flintknapper teaching a novice does not merely demonstrate but explains the purpose of shaping a tool in a particular way, even if the novice cannot immediately use the tool. This decoupling of representation from immediate action enables cumulative cultural learning and rich normativity by *explaining the general purpose detached from immediate response*.

Decoupled representations underpin *cumulative culture* as the progressive accumulation and refinement of knowledge and technology over generations. Sterelny traces the evolution of tool complexity, from simple stone flakes to composite tools and eventually modern technology, as evidence of this process. The capacity to hold abstract representations of tool function and design, detached from immediate use, allows humans to innovate by recombining existing knowledge in novel ways. Such “combinatorial creativity” can underpin normativity, as well.

Sterelny highlights the importance of *external scaffolding* represented by cultural artifacts and social institutions that support cognitive processes. Writing systems, calendars, and legal codes *externalize memory and planning*, reducing cognitive load and enabling abstract thought. The counterpart of this in modern social ontology might be Bratman’s theory of shared institutional agency based on planning (2022) which emphasizes planning as a core activity supporting shared agency. As Sterelny writes:

“External scaffolds offload cognitive demands, allowing humans to engage in forms of reasoning and planning that would be impossible unaided.” (Sterelny 2012b, 156).

These external cognitive tools are sometimes called *entanglements*: bidirectional dependencies between agents and artifacts (or environments). For example, the invention of written language not only externalized memory but also demanded new skills like literacy and infrastructures like schools. Moreover, as Hodder (2014) notes, artifacts “cannot reproduce on their own” as they require continuous maintenance, repair, and transmission across generations, binding humans to increasingly more complex webs of dependency further feeding back into cognitive evolution which Sterelny calls “downstream niche construction”.

Consider a smartphone: its use presupposes not just technical literacy but global supply chains for rare-earth minerals, energy grids, and software ecosystems. These entanglements drive cognitive evolution by demanding advanced planning, abstract reasoning, and cooperative problem-solving. Sterelny’s framework highlights how humans engineer their environments to scaffold cognition, but in doing so, they become “trapped” in self-created niches that impose new cognitive demands.

According to Bardone and Magnani (2007), a central mechanism in cognitive niche construction is *mimetic activity*, wherein individuals externalize fleeting thoughts and private ideas by representing them in the environment. This externalization is a form of disembodiment (**donald1991?**), creating representations independent of their authors and which can be manipulated, shared, and reinterpreted by others.

An illustrative example is found in cognitive paleoanthropology: cave paintings created by early humans served as external memory supports, *encoding information* about animals, locations, and vegetation relevant for hunting (Mithen 1988). These paintings acted as early information technologies, enabling the transmission and sharing of knowledge that exceeded the limits of individual memory.

Thus, cognitive niche construction through mimetic externalization creates *additional levels of representation* that augment internal cognition, facilitating complex problem solving, planning, and social coordination.

Extending this ecological-cognitive perspective to the social domain, social institutions can be understood as normative cognitive artifacts—complex, socially constructed systems of shared beliefs, norms, and rules that structure human interaction. Institutional theory, particularly the cultural-cognitive pillar (**scott2001?**), highlights that institutions are sustained by internalized symbolic representations that mediate between the external social world and individual cognition. Aoki (2011) also calls institutions “cognitive media” highlighting their intermediary position between individual cognition and social structure. internalized schemas encode normative expectations and shared meanings, enabling individuals to navigate social environments with reduced cognitive effort and uncertainty. This very much resembles Guala’s line of thought.

Institutions, thus, are not merely sets of rules or regulations but *embodied cognitive constructs* that emerge through and reinforce cognitive niche construction. They operate as *externalized, collectively maintained representations* that coordinate behavior, stabilize social order, and instantiate normative force. This view aligns with Bardone and Magnani’s emphasis on sharing representations: institutions distribute normative knowledge across individuals, embedding it in rituals, practices, language, and symbolic systems, thereby facilitating epistemic coordination on a societal scale. This reinforces out previous intuition about two representations with RiE—private and public one—where public one is encoded in others’ behavior and lives within the “X” in “if X do Y” and a private one guides an action of a certain agent.

Recently Stetelny has provided his own account of normativity (Sterelny 2021). Decoupled representations, according to Sterelny, are the *evolutionary precondition for normativity*. They allow agents to consider not just what *is* the case, but what *ought* to be done, and to represent social rules and expectations that are not reducible to immediate environmental cues.

Sterelny approaches the problem of normativity in terms of evolutionary puzzle of large-scale cooperation. While kin selection and reciprocal altruism explain some forms of cooperation, they are insufficient for the scale and stability of human social groups, which often involve unrelated individuals and *delayed or indirect reciprocity*. He frames the problem as follows: how did early humans solve the “free rider” problem—ensuring that individuals contribute to collective goods and do not exploit the cooperative efforts of others? The answer, he argues, lies in the evolution of normative cognition and the emergence of social contracts enforced by group—level monitoring and sanctioning.

Sterelny posits that normativity emerged through a co-evolutionary process involving:

* *Coalitional monitoring*: early humans lived in groups where individuals could observe and evaluate each other’s behavior. This transparency enabled the detection of free riders and norm violators. We can see it as an adaptation to a statistically unusual behavior;
* *Reputation and indirect reciprocity*: individuals who consistently contributed to group goals gained reputational capital, increasing their chances of receiving help in the future. Conversely, norm violators risked exclusion or punishment;
* *Collective sanctioning*: groups developed mechanisms for punishing defectors, ranging from gossip and ostracism to physical coercion. These sanctions were themselves subject to group norms, preventing the abuse of power. This presupposes that the sanctioning mechanisms themselves are “meta-normative” since they require social norms to be sustained.

Sterelny emphasizes that these processes depend on decoupled representations: agents must be able to represent rules, track reputations, and plan actions based on anticipated social consequences, not just immediate rewards or punishments. Sterelny writes:

“Normative cognition is not just about tracking what others do, but about learning and internalizing what one *ought* to do, and why. This requires a capacity for metarepresentation-for representing not just actions, but the rules that govern actions, and the reasons for those rules.” (Sterelny 2012b, 156).

We can suggest that agents not only extract a cue-action pattern in other agents and average over it to base their action choice on this higher-order cue, but *understand* how this particular pattern applies to this particular context. This requires abstracting a observed pattern from its immediate function in the context.

Sterelny’s “Pleistocene Social Contract” is not a literal agreement but an evolutionary equilibrium—a set of mutually reinforcing expectations and behaviors that stabilize cooperation within groups. This equilibrium is maintained by:

* *Mutual monitoring*: individuals observe each other’s behavior and share information about norm violations
* *Reputational feedback*: good standing in the group depends on adherence to norms; social rewards and penalties are distributed accordingly
* *Cultural transmission*: norms are transmitted across generations through teaching, imitation, and narrative.

This system is robust because it is self-enforcing: individuals have incentives to monitor, report, and punish norm violations, as their own reputations and access to group resources depend on the maintenance of cooperation.

Sterelny argues that the content of early human norms was shaped by the ecological and social challenges of the Pleistocene. Key domains include:

* *Food sharing*: norms developed to regulate the distribution of meat and other resources, balancing individual interests with group needs
* *Mating and kinship*: rules governing marriage, inheritance, and parental investment stabilized social relationships and reduced conflict
* *Conflict resolution*: mechanisms for mediating disputes and punishing aggression promoted group cohesion.

These domains remain central to human morality, suggesting deep evolutionary roots for contemporary normative systems.

Sterelny highlights the recursive relationship between norms and cognition: as groups develop more sophisticated normative systems, they create new cognitive demands like tracking complex reputational networks which in turn select for enhanced cognitive capacities. This feedback loop drives the cumulative evolution of both normativity and intelligence.

Sterelny’s model aligns with dual-inheritance theory, which emphasizes the co-evolution of genetic and cultural inheritance systems. This is what Gintis has put forward as “gene-culture co-evolution” (Gintis 2007a; Gintis and Helbing 2013). Sterelny argues that normativity is a product of both biological evolution (the capacity for decoupled, metarepresentational cognition) and cultural evolution (the accumulation and transmission of norms).

While Sterelny acknowledges the importance of language in the transmission and enforcement of norms, some theorists argue that he underplays its transformative impact. The emergence of symbolic communication may have been a crucial catalyst for the evolution of explicit, abstract norms.

Sterelny (2012a) also reflects on the evolutionary transition from fitness to utility maximization which is relevant in modeling the transition from ontic to epistemic correlation. He compares three models of human agency—instrumental rationality of utility maximization, instrumental rationality of strong reciprocation and evolutionary rationality of fitness optimization—and puts forward an argument of model pluralism due to their diachronic relation in evolutionary history.

As Sterelny notes, behavioral ecologists model agents as fitness maximizers and this works well in small scale. However, when social complexity increases and stratification and division of labour emerge, it is hardly possible to model agents as fitness maximizers, for there is no direct connection between environment and adaptive behavior.

The distinction between instrumental and evolutionary models of agency, according to Sterelny, lies in the sources of cultural information which helps to make decisions. For instance, in the hunter-gatherer social worlds information about other agents is freely available as a byproduct of established cooperation in small groups. As social groups grow larger, the source of cultural information ceases to be immediate, becomes mediated by others’ opinion and reputation.

With the change of the source of cultural information, agents don’t cease to be affected by natural selection and remain adaptable to the environment. However, population-level mechanisms change and this loosens the influence of fitness to individual decision making. As a result, proximate motivation is less tuned to adaptive ends of fitness.

Sterelny puts forward a hypothesis that with the rise of large-scale social worlds in the Pleistocene, there has been a significant decline in individual-level heritability of cultural traits, for offspring do not more resemble their parents informationally and ideologically due to the abundance of cultural information sources.

As the number of cultural information sources grows, the informational load on adaptive action increases. This induces development of symbolic systems like language and pictorial representation, expanded folk psychology that routinely employs joint planning and social norms. As this happens, individual decision-making does not longer reliably track fitness, and utility decouples from fitness which probably parallels the transition from ontic correlation of strategies to epistemic one.

Sterelny’s account is further nuanced by his endorsement of *function perspectivalism*, the view that biological functions are not intrinsic, fixed properties but are ascribed relative to explanatory interests and perspectives of agents.

This perspective contrasts with more rigid ontic accounts that treat functions as objective, mind-independent facts. Function perspectivalism allows Sterelny to accommodate the variability and context-dependence of cognitive functions, including decoupled representations.

For example, the function of a stone tool as a cutting device depends on the intentions and goals of its user, not on any intrinsic property of the stone itself. Similarly, the function of a decoupled representation is tied to its role in enabling flexible, goal-directed behavior within a given ecological and social context.

This stance has important philosophical consequences. It emphasizes the pragmatic and context-sensitive nature of cognitive functions and aligns with Sterelny’s broader evolutionary framework, where cognitive capacities evolve in response to shifting selective pressures and environmental challenges.

Sterelny’s theory has not gone unchallenged. One notable critique comes from (**christensen2010?**), who argues that many mammals exhibit forms of decoupling, such as food caching and future-oriented behavior, suggesting that decoupled representation is more widespread than Sterelny allows. Sterelny responds by refining his criteria, emphasizing that true decoupling requires *open-ended flexibility* as a capacity to generate novel, context-sensitive behaviors beyond innate repertoires. While squirrels cache nuts, their behavior is largely seasonal and stereotyped, lacking the strategic depth and flexibility characteristic of human planning. Sterelny clarifies:

“Decoupled representation is not merely the capacity to anticipate future states but the ability to manipulate internal representations *in novel ways*, supporting innovation and cumulative culture.” (Sterelny 2012b, 96).

Thus decoupled representation is *generative* and *flexible* just as Guala suggests, but it is nevertheless insufficient for explaining normativity and institutional stability yet.

## Decoupled representation is insufficient for explanation of normativity in RiE

As we have seen, decoupled representation is a cognitive capacity for tracking environment with context-sensitive, open-ended and flexible behavioral responses which overcome genetically “hard-wired” strategies as in most animals. And although decoupled representation allows for cumulative culture and downstream cognitive niche construction, it does not explain or account for normativity by itself as Guala and Hindriks would want.

According to Guala, decoupled representation enables humans to represent equilibria as meaningful rules and to coordinate flexibly on arbitrary correlated equilibria, thereby explaining the emergence and stability of social norms and institutions. However, a closer examination reveals that Guala’s argument is *internally inconsistent*. The core problem lies in the equivocation over what decoupled representation actually explains.

On the one hand, Guala attributes to it the power to ground normativity, or the “oughtness” and deontic force of social rules. On the other hand, the same cognitive capacity is acknowledged to be present, at least in rudimentary form, in non-human animals like chimpanzees exhibiting conventions and correlated behaviors that nevertheless lack genuine normativity as “second-order” beliefs. This conflation blurs the critical distinction between *coordination based on behavioral regularities* which I called *ontic correlation* and *normative coordination grounded in shared beliefs and motivations* which I called *epistemic correlation*. By not clearly differentiating these types, Guala’s theory risks treating normativity as an add-on extension of animal conventions, which undermines the explanatory distinctiveness of human social institutions (because animals would have them, as well, but they do not).

Beyond inconsistency, the argument is also *unsound* because it rests on a premise that decoupled representation suffices to explain normativity. And this premise is both philosophically and empirically contestable.

The core feature of RiE is using rules (social norms) as correlation devices in coordination games, which implies the pre-existince of norms to coordinate upon. However, although Sterelny’s theory of decoupled representation could bridge the divide between *ontic* and *epistemic* correlation of agents’ strategies by providing a mechanism for progressive inference of latent structure of the world over the course of evolution in repeated games, it still does not explain normativity and “following rules”, in particular. There is a gap between decoupled representation as a capacity to coordinate action on many representations (which is also questionable, for it resembles robust tracking rather than genuine decoupled representation) and intentional compliance to and creation of meaningful rules.

For Guala’s argument, decoupled representation is a necessary but not sufficient condition for explaining normativity and the stability of social institutions in the sense Sterelny envisions. While it underpins the cognitive flexibility needed for rule-following and the adoption of correlated equilibria (as Guala claims), the full explanation of normativity and institutional stability requires mechanisms for:

* norm enforcement and sanctioning
* internalization of norms as oughts and not just as “if-then” rules (akin to Bicchieri’s normative vs. empirical expectations)
* cultural scaffolding and learning, which embed rules in social practices and artifacts which *encode rules* like wedding ritual or traffic lights.

While decoupled representation indeed enables the cognitive possibility of rule-following and equilibrium selection, Sterelny’s own account shows that genuine normativity and institutional stability might arise only when these capacities are *integrated* with social, cultural, and motivational mechanisms that bind individuals to shared, enforceable expectations. Decoupled representation accounts for the cognitive possibility of representing abstract rules and engaging in flexible coordination, but it does not by itself explain why agents are motivated to follow rules or why rules carry prescriptive authority.

Normativity involves more than representation and requires internalized motivations, social enforcement mechanisms, and cultural scaffolding that give rules their binding character. Without incorporating these additional causal and motivational mechanisms, the premise that decoupled representation alone grounds rule-following and normativity is incomplete, rendering the overall argument less sound.

To fully explain how decoupled representation is responsible for normativity in human coordination games, we need:

1. to explain how conditional epistemic agents capable of inferring and coordinating upon the latent structure of the world (referring to Sterelny’s gradient of representational capacities) evolve from non-epistemic cue-based Skyrmsian ones. In other words, we need to show, how *ontic* correlation of strategies based on interactional dependencies between agents and environment (like “if bigger than opponent, fight”) evolves into *epistemic* correlation based on shared beliefs so that these shared beliefs can function as “features of environment” (“if North to the river, graze the cattle”);
2. to explain how observable behavioral regularities acquire normative (or deontic) force for epistemic agents as an endogenous feature of environment backed up by the capacity of decoupled representation to capture, store and contextually adapt to regularities of the latent structure of environment.

In short, how epistemic conditionalization evolves and how cognitive niche scaffolds emergence of epistemic normativity as shared normative beliefs, which are central to RiE.

To clarify, Odling-Smee, Lala, and Feldman (2003) has defined niche construction as the process whereby organisms modify environmental states, thereby altering the selection pressures acting on themselves, their descendants and other species. This process generates *ecological inheritance*, a form of non-genetic inheritance where descendants inherit not only genes but also modified environments. For example, beavers build dams that transform aquatic ecosystems, affecting not only their own survival but also that of many other species. Similarly, earthworms alter soil composition, influencing nutrient cycling and plant growth. These examples illustrate that organisms are not mere passive recipients of environmental pressures but *agents of environmental change*, creating feedback loops that shape evolutionary dynamics.

Odling-Smee, Lala, and Feldman (2003) articulate several key components of niche construction:

* *Environmental modification*: organisms alter abiotic and biotic aspects of their environment
* *Selective feedback*: these modifications change selection pressures on the constructor and others
* *Ecological inheritance*: modified environments are transmitted across generations, influencing evolutionary trajectories
* *Evolutionary consequences*: niche construction can accelerate, decelerate, or redirect evolutionary change.

Humans exemplify niche constructors par excellence. Beyond altering physical environments through agriculture, architecture, and technology, humans construct *cognitive and cultural niches* just as those scrutinized by Sterelny (2003). These include language, social institutions, and symbolic systems that create new domains of interaction and learning. Such constructed niches provide scaffolding for cognitive development and cultural transmission, effectively *co-evolving with human cognition*.

The human cognitive niche is thus a product of both biological and cultural evolution, where cognitive capacities and constructed environments form a dynamic, reciprocal system.

Traditional evolutionary game theory assumes that the payoff matrix is fixed and exogenous. This assumption implies that the environment, including social, material, and institutional factors that shape payoffs, is static. Consequently, the theory explains how strategies evolve *given* a certain environment but does not explain how the environment itself changes in response to agent behavior.

This limits the explanatory scope, especially for social norms, because the “rules of the game” (payoffs, available strategies, information structures) are often products of human activity and social construction.

Niche construction theory challenges the exogeneity of the environment by modeling how agents modify their environment, which in turn affects the payoff structure. This allows accounting for feedback loops where:

1. agents change environments
2. environments change payoffs
3. payoffs influence strategy adoption
4. adopted strategies further modify the environment.

Consider a population playing a public goods game, where cooperation yields collective benefits but is costly to the individual. Suppose agents can invest resources to build punishment mechanisms that sanction defectors, thereby modifying the payoff matrix by increasing the cost of defection.

* Initially, the environment has no punishment, and defecting dominates
* Some agents invest in building a punishment institution, changing the environment to
* The modified payoff matrix now includes penalties for defection, for example in the form of -parameters (Crawford and Ostrom 1995)
* This shifts the evolutionary dynamics, making cooperation more advantageous
* Over time, the population evolves toward cooperation, stabilizing the institution.

This example illustrates how niche construction can endogenously modify payoffs and enable the evolution of cooperation norms which some researchers put forward (Vlerick 2019, 2016; Skyrms 2014a). However, *our case does not suppose any change in payoffs*, since agent efficiently play CE from the outset in Maynard Smith’s Hawk-Dove-Bourgeois. What does niche construction affect in our case, then? The answer is *the degree of the functional abstraction of a correlation device* which corresponds to the level of cognitive sophistication of an agent.

## Correlated conventions with niche construction

Skyrms emphasizes that CE can be supported by *environmental cues or signals* that serve as correlation devices. For example, a traffic light’s color signals when to drive, enabling drivers to coordinate without direct communication with each other. These signals are part of the constructed social environment and play a crucial role in stabilizing conventions. The rules are thus *encoded* in these artifacts.

Niche construction theory complements Skyrms’s account by explaining how these environmental signals and cues are not fixed but *constructed and maintained by agents* themselves. The social environment, including traffic lights, linguistic conventions, or handshake rituals, is a product of collective activity that modifies coordination possibilities.

Herrmann and Skyrms (2021) model endogenous *invention* and *evolution* of correlated conventions. They integrate a Hawk–Dove game with a Schreiber urn reinforcement process (**schreiber2001?**) and an innovation mechanism derived from Hoppe’s urn model (**hoppe1984?**), thereby modeling the evolutionary dynamics of strategy invention, propagation, and stabilization. The model addresses the central question of how agents in a symmetric setting can converge on a stable, asymmetric behavioral convention *without external coordination signal* or “choreographer”. Their model shows that:

* *Invention* involves discovering or creating new environmental asymmetries or signals (via mutation)
* *Evolution* involves the population dynamics that favor coordination on particular environmental asymmetries.

The two-player Hawk–Dove game models a resource conflict in which mutual aggression is costly, mutual passivity is inefficient, and asymmetric behavior is optimal. Let each agent choose an action . The payoff matrix is defined as follows:

In Herrmann and Skyrms’s model, the value of the resource is , and the cost of conflict is , so the Hawk–Hawk outcome yields payoff 0 for both, and Dove–Dove yields 2.

Each agent is endowed with a vector of neutral traits like “size”, “brightness”, etc:

These traits are payoff-irrelevant and drawn independently from a uniform distribution on at the time of the agent’s birth. Trait values serve as potential *asymmetry generators* when agents adopt strategies that condition behavior on them.

The strategy space includes:

* *Unconditional strategies*: or
* *Conditional strategies* (or *conventions*) of the form:

where is the index of the trait used, and the comparison is either “greater than” or “less than”. For example:

“Play Hawk if my trait *x* is greater than (or less than) my opponent’s; otherwise play Dove”.

These strategies instantiate candidate *correlation devices* using arbitrary traits to resolve role assignment in an otherwise symmetric game.

The population is modeled as a Schreiber urn containing balls representing agents. Each ball encodes a strategy and a fresh draw of a trait vector .

In each generation:

* A number of agents are *sampled uniformly at random* from the urn and paired for interaction
* After playing the Hawk–Dove game, agents receive payoffs
* Payoff determines reproductive success: for each point of payoff, the agent’s strategy is copied into the urn with new, independently drawn traits
* A *mutator ball* is always present in the urn. When drawn, it *introduces a new conditional strategy*, based on a previously unused trait dimension and polarity (greater-than or less-than).

This mechanism follows *Hoppe’s model of innovation*, ensuring a continual possibility of *strategic invention*. It works like this: the urn starts a diverse collection of colored balls, including a black ‘mutator ball’. In each turn, a ball is drawn at random. If a color other than black is selected, it’s replaced in the urn, and a matching duplicate is added. But if the black ‘mutator ball’ is drawn, it’s also replaced and a new, previously unseen color is introduced. This process can continue without limit, resulting in an urn containing an infinite number of colors. These potentially infinite differently colored balls represent agent traits which are essentially *potential environmental asymmetries*.

Over time, conditional strategies that coordinate well on asymmetric trait values can *invade and outcompete* unconditional strategies. A particularly successful convention might be, for example:

“Play Hawk if my trait 2 is greater than my opponent’s, else play Dove.”

If most agents adopt this strategy, and trait 2 is randomly distributed, this induces a stable *correlated equilibrium* without any exogenous signal. The population effectively *constructs an asymmetry* out of a payoff-neutral trait. The model shows that:

* *Pure conventions* (all agents using the same trait and polarity) are *evolutionarily stable strategies (ESS)*.
* *Unconditional strategies* or mixed conventions are generally *unstable* under evolutionary pressure.
* The *timing of invention* matters: early-invented traits have a higher chance of becoming population-wide conventions due to *path dependence*.

Formally, model contains:

1. *Pairwise interactions in Hawk-Dove*
   * Each agent has traits randomly so that , where -th is a trait value for -th agent. Each time a new agent is added to population, Nature assigns these numbers randomly
   * Each pair plays the Hawk–Dove game, with payoffs computed based on actions , .

* *Reinforcement (reproduction)*:
  + For each unit of payoff , a new copy of agent ’s strategy is added to the urn, assigned a new independent trait vector .
* *Mutation and invention (Hoppe-style mutator)*:
  + The urn contains a persistent **mutator ball** . If selected during pairing, it invents a new conditional strategy not yet present in using an unused trait index and polarity.
  + This new strategy is added to and placed into the urn with a new trait vector.

The population starts in a state in which the only strategies agents play are the un- conditional ‘always hawk’ or ‘always dove’. The dynamics proceed similarly to the Schreiber urn model. At each time step a fixed proportion of agents are drawn from the population and paired off at random. Each pair plays the hawk–dove game, and replicate according to their payoffs.

The main difference from the Schreiber urn model occurs when a mutator ball is drawn.19 In this case a mutant is added to the population. There are two mutator balls, one of which we will call the big mutator ball and the other the small mutator ball. The ith time the big mutator ball is drawn a mutant with the conditional strategy ‘hawk if my ith property is larger than my opponent’s ith property, dove otherwise’ is added to the population. Similarly, the ith time the small mutator ball is drawn a mutant with the strategy ‘hawk if my ith property is smaller than my opponent’s ith property, dove otherwise’ is added to the population.20 Having both mutator balls allows us to use the model to explore how the contradictory strategies, Big i (Bi) and Small i (Si), interact. \*\*\*

The model demonstrates that *correlation devices need not be given*; they can *evolve from scratch* through mechanisms of invention, competition, and stabilization. By grounding the emergence of asymmetry in *trait-based comparisons*, the model offers a solution to the *problem of symmetry-breaking* in coordination games. Unlike classical correlated equilibrium frameworks like Aumann’s, this model does not assume a public signal or a referee. Instead, conventions emerge *organically* via decentralized dynamics.

This dynamic produces an evolutionary trajectory in which **novel conventions** can be invented and tested against the population, and successful ones are reinforced over time.

A key result of the model is that **pure conventions**, in which nearly all agents adopt the same conditional strategy based on the same trait and polarity, are **evolutionarily stable**. Formally, a strategy is an ESS if, for all mutant strategies , the expected payoff .

* In a population dominated by , coordination is consistent and efficient: each interaction resolves asymmetry using a shared comparison rule.
* Such conventions are **self-sustaining** under the urn dynamics and resist invasion by both unconditional strategies and alternative conventions.

By contrast:

* *Unconditional strategies* yield *lower average payoffs*, especially in mixed populations.
* *Mixed conventions* (with multiple traits or polarities in use) can persist temporarily but tend to be outcompeted by more coherent pure conventions.

The Herrmann–Skyrms model demonstrates that **correlated equilibria can evolve** without being externally imposed. Asymmetries that function as coordination signals need not be innate or environmental; they can be **invented**, **propagated**, and **stabilized** through selection on their functional effects. Trait comparisons become de facto **public signals** once a convention stabilizes.

This modeling approach supports a broader view in which **institutions and social roles emerge endogenously** through feedback between behavioral patterns and the structure of interaction. The model provides a baseline for exploring more sophisticated cognitive mechanisms, such as **flexible role representation**, **learning**, and **contextual generalization**—topics taken up in the next chapter.

Another important finding of Herrmann and Skyrms (2021) is that in populations where only Hawks and Doves have taken over the expected number of agents added to population after each round is 3 (two new and one “descendant” of a winner of the previous game), whereas in populations with a convention takeover it is 4 (because in CE both players have symmetric payoffs and both reproduce resulting in 4 new agents). It led Herrman and Skyrms to suggest that populations with established convention *would be larger* which might have consequences for group selection. This echoes Sterelny’s idea of the disentanglement of fitness and utility due to larger group sizes which, in its turn, promoted further informational complexity and decrease in individual trait heritability. This slightly suggests how having a convention affects further “ratcheting” of the evolution og cognition.

In this example, invention of new asymmetries serves as a form of niche construction, where agents randomly mutate to introduce new asymmetries and either get picked up by the rest of the population or not and if yes, it has both larger group size and further detachment of fitness from individual strategy choice.

While Guala’s framework explains how rules stabilize as equilibria, it does not fully specify how the *environmental and social conditions that sustain these equilibria emerge*. Niche construction theory fills this gap by showing how agents *actively build and maintain the material and symbolic environments* that embed rules as stable coordination devices.

* The constructed environment includes not only physical artifacts but also social norms, enforcement mechanisms, and shared representations.
* These environmental features modify payoffs and make certain equilibria focal and stable.
* Agents internalize rules through socialization within these constructed niches, acquiring the motivational and epistemic commitments Guala emphasizes.

A more direct treatment of informational and cognitive niche construction which we can incorporate into our model is due to Planer and Godfrey-Smith (2021), who generalize Godfrey-Smith’s model to explain coordination within and between organisms. After Skyrms (2010a), they group selection processes that can stabilize signaling and present them in order of increasing cognitive requirements:

* biological evolution
* reinforcement learning
* imitation of the successful
* rational choice.

After Godfrey-Smith, they discern state–act and act–act coordination. Between-agent coordination mostly functions as information transmission, whereas within-agent coordination works as routine combination and processing to yield new information which is useful in controlling behavior. In other words, the latter kind of coordination not only transmits information, but carries out computations. This fact, authors suggest, establishes co-evolutionary feedback loop between the within-agent and between-agent coordination processes. With the increasing organization of internal signaling, computation becomes more efficient, and it allows to represent more external information in terms of states of the world so that one can act upon them.

Sterelny (2017) critiques sender–receiver framework for its apparent inability to accommodate “mindreading” or recursive reasoning about others’ beliefs, for it requires a theory of mind, and the framework is not equipped to grasp it. However, Planer and Godfrey-Smith reply that sender–receiver framework is flexible enough to allow gradient in amount of cognitive resources required: it accommodates a spectrum of possibilities from those not requiring any meta-psychology or mindreading to strategically or interpretively sophisticated ones.

Although this “model” in the most appropriate for our case of substantiating Guala’s RiE with niche construction, it still has drawbacks. It explicitly appeals to the epistemic shift in agents’ cognitive capacities and accounts for feedback between internal computations, cognitive affordances and conventions’ strength, but it does not have a mechanism for detaching signals from fixed referents paving the way to decoupled representation and normativity.

| Model | Ontic Corr. | Epistemic Corr. | Decoupled Repres. | Normativity |
| --- | --- | --- | --- | --- |
| Skyrms (2004, 2010) | ✓ | × | × | × |
| Godfrey-Smith (2021) | ✓ | Partial | Partial | × |
| Guala (2016) | × (assumed) | ✓ | × (assumed) | ✓ |

To make the case for the *emergence of epistemic correaltion and normativity*, we still need:

1. Show how *ontic coordination* (mere behavioral fit) gives rise to *epistemic correlation* as coordinating on shared expectations and beliefs
2. Show how *shared representations* emerge *as stable artifacts* via niche construction
3. Model how agents begin to *represent rules* and act on them independently of immediate payoffs — *decoupled representations*
4. Show how such representations *constrain and enable* new equilibria — the key point in Guala’s “rules-in-equilibrium” theory.

That means that we need to:

* introduce a dynamic where *agents represent the game they’re in*, rather than just play it reactively
* embed niche construction as a *process that shapes the game itself*, making *rule-representation adaptive*.

After Sterelny, we will use niche construction as a main evolutionary mechanism for both emergence of epistemic conditional agents from non-epistemic ones and evolution of normativity with institutional scaffolds. Developmental plasticity will be a key feature of a baseline cognitive architecture. Next, we formulate the problem in terms of cognitive architectures of reinforcement learning and Bayesian updating, which will serve as a foil for our own model.

## Cognitive architectures of Skyrms’s and Guala’s agents and the evolutionary plausibility of the shift

Guala’s RiE can benefit from taking Sterelny’s theory of decoupled representation theory in its entirety seriously. It can be done by introducing the mechanism of cognitive niche construction as a bridge between ontic and epistemic correlation of strategies. As we have also distinguished between two agency models within RiE — Skyrms’s non-epistemic conditional and Guala’s epistemic conditional agents, their agents differ in cognitive architectures or underlying rationality assumptions. And niche construction theory can help them meet. Moreover, the very same niche construction mechanism can account for emergence of full-fledged normativity in the form of epistemic correlation of normaitve beliefs as in Guala’s “Grazing game” and payoff-modified games. This would require us to model these transitions explicitly, but to do that we need to reformulate the problem in terms of cognitive architectures.

Skyrms’s agents do not possess any shared beliefs and coordinate strategies using solely interactional dependencies and environmental asymmetries. They are cue-based detection systems, in Sterelny’s terms, meaning that they possess the simplest possible representational grade of tracking only genetically hard-wired environmental cues. Guala’s example of butterflies correlating strategies based on the order of occupying of a sun spot is an example of such agent and of an objective non-epistemic environmental dependency. From a modeling point of view, Skyrms’s agents are *simple reinforcement learners (RL)* (Erev and Roth 1998) who update strategies on the observed objective frequencies and not representations. They do not have any *beliefs* to update.

Guala’s agents do possess shared beliefs, but they are boundedly rational, use simpler heuristics and rules of thumb and do not engage into full-fledged Bayesian reasoning with cognitively demanding models like Vanderschraaf’s Dirichlet dynamics (which also produces endogenous CE). They coordinate on shared social norms as correlation devices, which are, importantly, encoded in the environment. For example, traffic lights, dried riverbeds and banknotes all, as I will argue after Sterelny and ecologists, *encode* and *scaffold* normative information in material base. Guala’s agents do update their beliefs, but they seem to base them on observed frequencies, which points to modeling them as hybrid RL-Bayes agents. I assume it from Guala’s appeal to repeated games and “learning from experience” which his agents do.

Formally, the core difference between the two types is in the capacity for inferring and acting upon the latent structure of the world, which lies at the backbone of Sterelny’s taxonomy of representational capacities: as the representational sophistication, according to Sterelny, *depends* on the capacity for accumulation and integration of various information, a mechanism for gradually learning to infer and use the latent structure of the world to increase one’s fitness is necessary. Robust tracking is an example: many birds detect predators by combining auditory, olfactory and visual cues. And, as some scholars note, such a capacity for integration requires a form of conditionalization, Bayesian or otherwise (Samir Okasha and Okasha 2013). However, there are varying views on imbuing agents with capacities for Bayesian updating, as it is considered a rather more cognitively demanding procedure (Guala 2019; Bain 2016; J. M. Alexander 2009).

Nevertheless, radical approaches like “active inference” (Friston 2010; Friston et al. 2013) provide a low-level view of perception, action and inference as a Bayesian inference, even in simplest organisms. They even purport to explain normativity mechanistically as surprise minimization and approximate Bayesian inference saving cognitive effort, but these mechanisms have mathematically and computationally complex descriptions, which would complicate our argument, so we will not use them here.

While both Skyrms’s and Guala’s agents can engage in correlation of strategies, the former ones do not possess a capacity for conditional *reasoning* and the latter do, for they explicitly use it to follow the rules. What we need here is an explanation of emergence of conditional reasoning first.

Overall, our argument depends on showing the evolution of Sterelny’s representational grades from non-epistemic cue-based agents. We can model it by employing Sterelny’s appeal to the environmental translucency, which is noisiness of environment which can potentially embed latent structure acting or coordinating on which can be evolutionarily advantageous and fitness-maximizing. It means that agents in the course of evolution learn to track, use and coordinate upon the latent structure of environment.

Samir Okasha (2015) shows that evolution of Bayesian conditionalization is evolutionarily plausible from a formal point of view. Bayesian updating, or conditionalization, prescribes that an agent’s prior probability distribution over possible states of nature should be revised to a posterior distribution upon receiving new evidence, in accordance with Bayes’ rule:

where denotes a possible state of nature and a possible signal or observation.

Philosophers have long debated whether Bayesian updating is a requirement of rationality. Okasha asks a different question: *Is Bayesian updating favored by natural selection?* That is, will organisms that update their beliefs in a Bayesian manner enjoy a selective advantage over those that use alternative updating rule?

Okasha formalizes the problem as follows:

* Let be the finite set of possible states of nature.
* Let be the finite set of possible actions available to the organism.
* The payoff from action in state is denoted , interpreted as expected reproductive fitness.
* The organism receives a signal from a finite set , probabilistically related to the true state.
* The organism’s **policy** is a rule that, for each signal , prescribes an action .

The organism’s expected fitness, given policy , is:

where is the prior probability of state , and is the probability of observing signal in state .

Okasha proves that the *evolutionarily optimal policy* that maximizes expected fitness is the policy that, for each signal , chooses the action maximizing the *conditional expected payoff*:

where the posterior probability is calculated using Bayes’ rule. Thus, the optimal policy is to act as if one were a Bayesian agent: update beliefs by conditionalization and choose the action with the highest expected fitness given those beliefs.

As an example, Okasha provides a simple foraging scenario. An organism must choose how to respond to possible predators like a snake, a leopard or none based on ambiguous environmental signals (like Sterelny’s translucency). The organism’s best action depends on the true state, but the state is not directly observable. By calculating the expected fitness for each action using Bayesian-updated probabilities, the organism can maximize its evolutionary success.

The key implication is that *natural selection can favor the evolution of Bayesian updating*, not because organisms consciously calculate probabilities, but because policies that approximate Bayesian rationality yield higher expected fitness. Over evolutionary time, such policies will outcompete alternatives.

Okasha’s result provides an *evolutionary justification for Bayesian rationality*: crucially, organisms that behave *“as if”* they are Bayesian updaters integrating prior knowledge and new evidence optimally will be favored by selection in environments where information is valuable for survival and reproduction.

Okasha notes that this conclusion depends on the assumption that *expected reproductive output is the sole determinant of evolutionary success*. In real biological systems, other factors like costs of information processing, environmental complexity may complicate the picture. Nonetheless, the argument establishes a strong link between evolutionary optimality and Bayesian rationality.

Okasha’s view of agents as “as if” Bayesian is consistent with Skyrms’s approach, where non-rational can produce outcomes which looks rational. It means that organisms might not actually be Bayesian from the point of view of their actual cognitive architectures. The challenge, therefore, is to show how *actual conditionalization* evolves from “as if”-conditionalization.

Tracking of and coordinating upon the latent environmental features can be modeled as conditional inference, where occurrence of one cue is assessed according to the occurrence of another one, thus establishing associations. This inference itself evolves from the interaction of agents with environment.

As Skyrms’s agents do not possess any rationality and beliefs, although they exhibit seemingly rational aggregate behavior, they are modeled either as evolutionary agents (phenotypes) or as simple learners (reinforcement learning agents) (Skyrms 2010a, 2014a). We will model them as RL agents. Guala’s agents are trickier to model as there is no straightforward way in the literature to represent them. Although they are conditional, epistemic and coordinate on shared beliefs, Guala portrays them as boundedly rational agents learning from observed empirical frequencies akin to RL-agents. And although they do not express full Bayesian rationality, they still have conditional probabilities on beliefs given certain signals, meaning that they can be modeled as Bayesian agents. This is why we will model them as hybrid RL-agents with Bayesian updating components.

We are now ready to set up our model of emergence of epistemic correlation from ontic one with niche construction which evolves conditionalization with decoupled representation, effectively solving Guala’s problem of “scope of actionable signals” as differentiator between animal and human conventions.

## Contributions of our model and implications for RiE

One of the core contributions of our model is the *operationalization* of the distinction between *ontic and epistemic CE* in coordination. Ontic correlation relies on reactive, probabilistic, frequentest cues while epistemic correlation depends on *conditional beliefs and representations* of roles, norms, and strategies.

This distinction reflects the dual roles of *cue detection* and *cue interpretation* in evolving social behavior. We draw a conceptual bridge between:

* *Ontic correlation*, grounded in evolutionary game-theoretic mechanisms like ESS (Maynard Smith, Skyrms), where coordination is stabilized through fitness landscapes
* *Epistemic correlation*, driven by *conditional reasoning*, role-representation, and agent models of others’ expectations, often associated with conditional reinforcement learning or Bayesian cognition.

Our model shows how agents can move along this ontic–epistemic gradient through interaction and environmental structuring, echoing Sterelny’s proposal of cognitive decoupling and supplementing it with dynamic niche construction and learning feedback.

This transition is significant: it suggests that *institutional regularities need not be “imposed” top-down via rational agreement*, but can *emerge from ecological scaffolding* that incrementally builds representational capacity.

Our model provides a mechanistic framework for understanding how equilibria in rules-in-equilibria emerge, stabilize, and persist. Crucially, we show that representation indeed can be a precondition for a normative equilibrium. In addition, the “scope of actionable signals” acquires its tangible non-metaphorical and mechanistic form. This “scope” is not quantitative, but qualitative: it is the scope of abstraction and generalization.

Institutions in our model arise as *distributed solutions to coordination problems*, stabilized through recurrent role-based asymmetries and cue regularities. Agents eventually “internalize” role-action mappings, leading to proto-deontic structure: certain actions become expected in particular contexts. This aligns with theories of institutions as *cognitive technologies* (Aoki 2011), but improves on prior models by:

* Grounding the emergence of representation in low-level reinforcement learning and ecological constraints
* Avoiding assumptions of centralized cognition or external enforcement
* Demonstrating how salience and role perception jointly scaffold proto-deontic norms.

In this light, our model supports a *gradualist view of normativity*: norms emerge through ecological dynamic feedback and structural learning, not through linguistic or conceptual innovation.

Guala argues that CE is “too loose” to explain the stability of human social institutions, claiming it fails to explain how human coordination is different. He appeals to Sterelny’s decoupled representations as the distinctive cognitive feature enabling humans to maintain institutional equilibria. However, I find a conceptual inconsistency in Guala’s framing: he happens to critique evolutionary models like ESS for failing to explain institutional resilience, yet simultaneously motivates the need for decoupling via animal behavior. This *conflates evolutionary and epistemic models* of agency.

I respond by introducing a new distinction:

* *Ontic correlation*: evolutionary, reactive, cue-based coordination as seen in animal behavior which coordinate on a small mostly genetically-wired strategies
* *Epistemic correlation*: conditional, decoupled, role-sensitive belief-based coordination.

Our model shows how *epistemic correlation can evolve from ontic correlation* under certain environmental pressures. This avoids the need to posit a cognitive leap between animal and human coordination. Instead, it provides a *mechanistic bridge*, where environmental feedback and niche construction foster increasingly decoupled role representations.

Moreover, I demonstrate that *decoupling alone is insufficient* for institutional stability. Stability arises not from cognitive decoupling per se, but from *the embedding of such representations in structured environmental feedback loops*, leading to stable cue-role-action triads. It might even be extended further to facilitate genuine normativity (with violation-detection), but I will leave it for future research.

This insight complements and refines Guala’s account: institutions are not simply equilibria of expectations, but *emergent systems of role-conditioned strategies shaped by ecological feedback*, thereby grounding social rules in ontological regularities.

Overall, Guala’s theory of institutions as rules-in-equilibrium was designed to reconcile two historically divergent strands in institutional ontology: the Searlean constructivist view (institutions as “constitutive rules” of social facts) and the *Humean/game-theoretic approach* (institutions as equilibria in strategic interaction). In Guala’s synthetic view, rules are not causes of behavior but *labels* (like “money”) we assign to self-sustaining patterns of interaction that solve recurrent coordination problems. My model supports the basic intuition of Guala’s synthesis that institutions are equilibria sustained by collective expectations but substantially revises its underlying mechanism and ontological commitments. Specifically, I make three critical updates:

### From cognitive labeling to generative mechanisms

Guala interprets rules as *descriptions* of conventions that stabilize through shared expectations. This view is largely representational and epistemic: rules exist *in virtue of being recognized* as rules. Searle, in contrast, holds that constitutive rules *create the very possibility* of institutional facts (“X counts as Y in C”). Our model *bridges this divide* by showing how such representations themselves can *emerge endogenously* from non-representational dynamics. Rules are not *imposed* from above via constitutive declarations, nor merely *recognized* from below through cognitive labeling. Instead, they *emerge through repeated interaction*, shaped by environmental feedback and evolutionary learning, and only *later* acquire representational form.

This leads us to a *generative view*: institutional facts arise through the evolution of role-action mappings embedded in ecological contexts. The “rule” is first an *informational regularity* between cues and strategies, later internalized as a *proto-deontic expectation*. In this sense, our model offers a *mechanistic ontogeny of rule formation* and not merely a formal description of equilibrium.

### Naturalizing the constitutive/regulative divide

Searle’s famous distinction holds that:

* *Constitutive rules* create new forms of behavior (“the rules of chess constitute the game”)
* *Regulative rules* guide behavior that exists independently (“drive on the right side of the road”).

Guala collapses this distinction by arguing that both kinds of rules are *epistemically dependent* on equilibrium behavior: the rule does not cause the institution, but tracks stable mutual expectations. However, this move risks *flattening the causal-explanatory structure* of institutional emergence.

Our model proposes a hopefully deeper reinterpretation. Constitutive rules can be understood not as declarative acts of fiat, but as the *retrospective codification of emergent role-regularities*. In our simulations, stable coordination strategies like owner plays Dove or intruder plays Hawk eventually yield proto-deontic expectations (“if I am uncertain and other seem to play Dove, I play Dove, too”). This is not an “ought” yet, but functionally it is quite close. These role-specific behaviors are *institutionally generative* in functions as *they constitute the institution*, even if no explicit rule is initially declared.

Conversely, regulative rules emerge as *meta-strategic refinements*: cultural, moral, or legal overlays that fine-tune coordination outcomes without altering the underlying roles (“don’t trespass” reinforces the ownership role distinction). This framework *reconstructs the constitutive/regulative divide* not as a logical distinction, but as a *temporal-functional gradient* in norm emergence:

* *Early-stage* emergent regularities are *functionally* (rather than semantically) constitutive of the institution
* *Late-stage* codifications: *regulative*, reinforcing or formalizing expectations.

This reframing allows us to *naturalize the Searlean schema*: “X counts as Y in C” is not a primitive speech act but an emergent generalization of stable cue-role-action patterns. The performative dimension remains, but it *builds on representational structures* that evolved through coordination dynamics.

### Environmental realism and the ontology of rules

One of the lingering problems in both Guala’s and Searle’s accounts is the *ontological status* of institutional rules. Are they socially constructed, or do they reflect something objective?

Our model advances a *naturalized realism* about institutional kinds. Rules are not arbitrary constructs or purely mental states but *real patterns*[[33]](#footnote-91) *in agent-environment interaction*, stabilized by ecological feedback and learning. This aligns with structual realism with its emphasis on multi-level real patterns (Ladyman et al. 2007; Beni 2017), Boyd’s homeostatic property cluster theory (R. Boyd 1991, 2021) and Psillos’ causal realism (Psillos 1999): rules (as role-conditioned conventions) are *causally efficacious*, *stable under intervention*, and *explanatorily indispensable*.

Furthermore, by grounding coordination in *ontic correlations* we secure an essentially *non-representational substrate* for institutional facts. The institution of “ownership,” for instance, arises from role asymmetries and behavioral regularities that can be described independently of any particular agent’s belief.

This perspective allows us to:

* Preserve the explanatory power of Searlean institutional facts without relying on constitutive declarations
* Deflate Guala’s anti-realist leanings by grounding institutions in *robust coordination mechanisms* rather than descriptive conventions
* Reorient debates about social construction: what is “constructed” is not the *existence* of social rules, but their *representational overlay*.

In terms if Epstein (2015), who distinguished grounds and anchors in social ontology (instantiation conditions of a social fact and its mechanism), we naturalize both. Ecological emergence of role-action-observation triads provides a causal mechanism for setting up a higher-order institutional fact “this is private property” (although, not directly, but through Hindriks-Guala reduction of constitutive to regulative rules) and the conditions for this fact to be true are not “socially constructed” in the sense of being conjured up ex nihilo but are grounded statistical regularities of the environment. Overall, this is essentially social ontology beyond the “Standard Model”.

One more important thing is Guala’s serious treatment of the notion of *kinds* in relation to social ontology. After Hacking (1999), he distinguishes between *indifferent kinds* in the natural sciences that are unaffected by how they are classified and *interactive kinds* that change in response to classification. Social (or institutional) kinds, in Guala’s account, are interactive: the act of labeling, identifying, or representing them causally contributes to their ongoing stability or transformation.

Our model captures this insight and pushes it further. As conventions evolve, agents increasingly rely on role-based reasoning. The assignment of a role (“owner”) is not just a neutral description; it shapes expectations and strategic behavior. Once a convention has stabilized around a representational pattern, that representation becomes *causally constitutive* of the interaction itself. It is no longer just a belief about roles; it is part of the mechanism that sustains coordination. Importantly, I do not confine institutional emergence only to roles as cues. In line with Skyrms’s “correlated conventions”, I believe any interactional asymmetry can be an initial cue for agents and role is what specifically suits Guala’s emphasis on the institution of private property.

Moreover, our simulations reveal how conventions shift from being *cue-driven* (and thus more indifferent) to being *representation-driven* (and thus more interactive). This transition mirrors Guala’s notion of interactive kinds becoming entrenched. When agents rely on shared expectations and internalized roles, the social kind like “ownership” gains stability through mutual recognition, even though it still remains ontologically dependent on representational practices.

Thus, the model bridges the gap between causal and constitutive dependence in Guala’s account: representations do not merely describe the convention but participate in its causal maintenance. This provides a naturalistic foundation for Guala’s theory: social kinds are interactive not by fiat but because *feedback loops integrate representation and behavior*, generating kinds that are both real and reflexively sustained.

By updating Guala’s rules-in-equilibrium theory, our model contributes a *rich, multi-level ontology* of institutions:

* At the *ecological level*, institutions are cue-action-observation regularities shaped by selection pressures
* At the *cognitive level*, institutions are representational models used by agents to navigate coordination problems (very much akin to Aoki’s “cognitive interfaces” (Aoki 2011))
* At the *cultural level*, institutions are normatively-driven conventions, embedded in language, law, and morality.

Each layer *emerges from the lower level*, but maintains partial autonomy, providing a bridge between ontological realism and constructivist accounts.

This framework lends support to *scientific realism about social kinds*, as advocated by Chakravartty (2007) and R. N. Boyd (1983), while also explaining how *normative representations can evolve naturally*. It invites a reappraisal of Searlean institutional ontology through the lens of ecological naturalism and offers a new foundation for explaining both the emergence and stability of social institutions.

## Conclusion: implications for social ontology, social science and naturalistic metaphysics

What does a broadly ecological explanation of emergence of epistemic agency and normativity give to social ontology? This thesis began by identifying a conceptual tension in Guala’s influential theory of rules-in-equilibria. Guala argues that correlated equilibrium (CE) is too permissive to explain the stability of social institutions, invoking the distinction between animal coordination and institutional normativity. However, his critique of CE inadvertently blends two incompatible modeling perspectives: evolutionary (dynamic, model-free learning) and strategic (static, belief-based rationality). By referencing Maynard Smith’s evolutionarily stable strategies (ESS) alongside concerns about epistemic commitment, Guala presumes that institutional stability must be underpinned by cognitive capacities for decoupled representation, which is, as I have shown, is insufficient.

This realization led to the central innovation of our model: a new distinction between *ontic* and *epistemic correlation*, corresponding respectively to *evolutionary-reactive* and *representational-conditional* forms of agency. Ontic correlation emerges from repeated co-adaptive interactions and niche-structured cue use. It uses environmental asymmetries to produce Skyrmsian “correlated conventions.”, which I claim to be a feature of the ecology of interactions, not the beliefs of agents. Epistemic correlation, by contrast, involves the internalization of cue expectations and conditional reasoning over context-sensitive cues.

Our model demonstrates how epistemic correlation does not need to be presupposed to explain stable institutional behavior. Instead, it can *emerge from ontic structures* through processes of cue learning, environmental scaffolding, and reinforcement dynamics. This is achieved by extending Sterelny’s theory of decoupled representation in an original direction: rather than treating decoupling as a cognitive leap that divides animals from humans, we integrate it within a *gradual ecological-evolutionary model*. In our system, agents begin as detection-based cue followers (ontic learners) and evolve through increasingly structured conventions to become epistemic agents capable of conditional inference. The transition is driven by coordination pressure, cue translucency, and the need for higher-order environmental mediation.

This ecological enrichment of Sterelny’s view helps overcome a key limitation in Guala’s approach. While Guala rightly stresses that social institutions require more than animal-like correlation, he attributes this surplus to a shift in cognitive architecture alone. I have shown that *proto-normativity responsible for institutional stability emerges not just from internal representation, but from the interaction between cognitive types and cue-structured environments*. Representational content, in our model, is scaffolded by ecological regularities and evolutionary selection and not merely posited as a precondition for normativity.

This reframing allows us to revisit the role of conventions in social ontology. Recalling O’Connor’s typology of conventions (O’Connor 2019) (functional/arbitrary, non-normative/normative), our model shows how norm systems may evolve from *functional, non-normative conventions* to *arbitrary, normative ones*. The early stages of cue-based coordination are driven by ecological functionality: agents adapt to salient features of the environment. But as the system stabilizes and decoupled cognition evolves, conventions can drift from being ecologically necessary to being *symbolically reinforced and socially normative*.

This shift from functional to arbitrary conventions, and from non-normative to normative forms, is a central empirical finding of our model. It demonstrates how normativity can emerge from coordination pressures alone without needing top-down authority or intrinsic moral content.

Moreover, the emergence of (proto)-normative structure is correlated with the loss of ecological transparency: *the more decoupled the convention, the more crucial it is for agents to internalize role expectations and enforcement mechanisms*.

Crucially, our model does not posit this transition as a mystery. Instead, it provides a mechanism for the *gradual emergence of normativity from strategic behavior*, shaped by ecological feedback, cognitive evolution, and reinforcement dynamics. In this respect, our approach helps close the gap between *causal explanation in the natural sciences* and *interpretive explanation in the social sciences*.

By demonstrating how social conventions can be modeled as *dynamically stabilized, ontically constrained structures*, we have advanced a new vision for *social scientific metaphysics*. Social kinds are not mere projections or fictions; they are *products of ecological, cognitive, and evolutionary regularities*, much like species, ecosystems, or morphogenetic patterns. When sociology models these structures as emerging from the interplay of agents, cues, and environments, it positions itself as a *metaphysically serious science*, grounded in ontic structures that are amenable to explanation, prediction, and intervention.

The evolutionary model presented above, while designed as a simplified simulation of strategic interaction and cognitive development, also carries deep implications for how we conceptualize social kinds, the structure of social ontology, and the epistemological commitments of social science. In particular, it raises a fundamental question: *Are social conventions and norms natural kinds?* If so, what constraints does this place on sociological theorizing?

### Conventions as natural kinds

In our model, conventions (e.g., “owners play Hawk, intruders play Dove”) emerge through a process of reinforcement and selection that is both mechanistic and empirical. These conventions stabilize because they confer fitness advantages to individuals who conform to them. This process meets several core criteria for natural kinds in a scientific ontology:

* **Causal Role**: The conventions in the model do real explanatory work. They causally mediate the behavior of agents and determine success or failure in coordination, making them indispensable to any predictive or explanatory account of the system.
* **Projectibility and Induction**: Once a convention stabilizes in the population, it supports inductive generalizations: agents behave in consistent ways in new interactions, and the convention reliably predicts both action and outcome.
* **Integration with a Causal Theory**: The conventions are not ad hoc descriptions but arise within a broader explanatory framework—evolutionary game theory and reinforcement learning. They are integrated into a theoretical structure with empirical content and predictive power.

Taken together, these features justify treating conventions as **natural kinds** in the sense required by a scientific ontology. They are repeatable, causally efficacious, and explanatory entities that emerge from the dynamics of strategic interaction and adaptation. In this respect, they are more than mere social constructs or contingent patterns—they are *evolved regularities* in behavior and cognition, subject to evolutionary explanation.

### Ontological constraints and legitimate social kinds

Accepting this view entails a significant ontological constraint on social science (and ontology, although it sounds pleonastic). If conventions are natural kinds and form the foundational causal structure of social interaction, then many higher-level sociological entities such as roles, statuses, institutions and classes can be understood as *derivative* or *compositional kinds*, whose legitimacy as scientific objects depends on their grounding in evolved normative conventions.

This implies a stricter epistemological foundation for social science:

* *Legitimate social kinds*: only those entities that can be shown to depend on, emerge from, or be reducible to causal conventions should be treated as legitimate explanatory objects. Social “roles” make sense insofar as they encode stable conventions about behavior; “institutions” are just structured aggregations of such roles; “social classes” may be seen as statistical clusters of convention-governed roles and access to resource-contingent strategies;
* *Exclusion of unfounded entities*: categories not anchored in evolved conventions that cannot be shown to exert causal force or guide behavior in stable, predictable ways should probably be excluded from theoretical consideration. This serves as a corrective to sociological theorizing that relies on reified and somtimes ambiguous categories like “culture” or “ideology” without clear causal specification. This supports both Gintis’s view of unifying social science with evolutionary grounded and computationally tractable mechanisms (Gintis 2007a; Gintis and Helbing 2013) and the middle-range theoretic optics of analytical sociology striving to provide causal mechanisms (Hedström and Bearman 2009);
* *Grounding normativity*: social norms, often treated as abstract or purely normative constructs, are here reinterpreted as the emergent result of evolved coordination mechanisms. Their normativity derives from their fitness-enhancing role, not from moral or legal fiat. In this way, normativity itself becomes a *naturalized* object of study.

Thus, the model does not merely describe how conventions emerge but places them at the *center of social ontology* and elevates them to a primary explanatory category. All other social objects, insofar as they are legitimate, are reducible or relatable to these evolved, causal conventions. as Dennetian “real patterns” which have their own existence.

### Bridging evolutionary theory and social ontology

This modeling framework supports a broader philosophical and methodological synthesis: a *naturalistic integration of evolutionary and especially ecological theory with the foundations of social ontology*. Such an integration would yield multiple benefits:

* *Unified explanatory framework*: by linking social coordination, cognition, and convention formation within a single evolutionary framework, we gain a powerful, interdisciplinary toolkit for explaining how social structures arise and stabilize;
* *causal clarity*: evolutionary models force us to specify mechanisms of transmission, selection, and stabilization. This guards against vague or metaphorical explanations that have historically inhabited parts of social science;
* *constraint-based epistemology*: an evolutionary approach *constrains the space of possible explanations*. not every socially observed pattern can be treated as causally or ontologically significant but only those consistent with plausible mechanisms of emergence, adaptation, and transmission;
* *Cumulative scientific progress*: naturalizing social ontology allows for better integration of findings across disciplines from behavioral economics to cognitive science to anthropology directly following both Guala’s and Gintis’s ambitious unificationist projects. The evolutionary modeling of conventions creates a shared basis for cumulative theory building.

The broader implication of this work is a further expansion and deepening of a *naturalistic social ontology*. Rather than treating the social world as an interpretive domain governed by historically contingent categories, we advocate treating it as an evolved, structured system whose properties can be uncovered through causal modeling and evolutionary reasoning.

This turn does not eliminate interpretive or normative questions but grounds them in a somewhat firmer epistemological foundation. It allows us to ask not just *which social meanings exist*, but *why*, *how they emerged*, and *under what conditions they persist or fail*.

Conventions, then, are not just behavioral regularities or semantic mappings. They are the *atoms of social ontology*: evolved, selected, and causally powerful kinds upon which all higher social structure rests.

### Conventions, scientific realism, and social scientific metaphysics

The evolutionary model of convention formation developed above invites a broader yet modest re-evaluation of the metaphysical foundations of social science. By treating conventions and social norms as causal structures evolved with the single dynamic with explanatory force, we reposition them as legitimate natural kinds within a realist scientific ontology. This reframes the longstanding question of whether the social sciences can aspire to the same metaphysical depth and explanatory robustness as the natural sciences.

Scientific realism, broadly construed, holds that successful scientific theories are (approximately) true representations of a mind-independent reality, and that we ought to believe in the reality of the entities and structures posited by our best theories (R. N. Boyd 1983; Psillos 1999). On this view, commitment to the reality of quarks, genes, or tectonic plates is justified by their indispensable explanatory role in empirically successful theories.

Our model extends this commitment to *social conventions and norms*, not as artifacts of human intention or interpretation (at least partially), but as *evolved, causally efficacious structures*. These structures:

* *Exert explanatory force*: They are not mere descriptions of patterns but explain why coordination succeeds, why particular strategies stabilize, and why certain social forms persist;
* *Possess counterfactual robustness*: Intervening on cue availability, agent cognitive type, or environmental structure changes the presence and form of conventions demonstrating their integration into causal networks. This goes in line with seeing equilibrium explanations as interventionalist (Woodward 2005), where changes the control parameters of the model produces causal effects (Sperry-Taylor 2021);
* *Support inferential roles*: Once conventions emerge, they serve as a basis for prediction, intervention, and explanation, quite like biological or physical entities.

From a Boyd–style homeostatic property cluster view of natural kinds, conventions are kinds not by virtue of essential properties and ncecessity but because of the clustering of properties maintained by causal mechanisms: reinforcement learning, environmental feedback and social selection. Their kindhood is sustained through adaptive processes that stabilize their features over time and across contexts. It goes in line with Guala’s note that institutional kinds are stable, yet not necessary.

Martínez (2020) gives a further refinement to natural kinds as *synergic kinds*, claiming that it is not the redundancy and co-occurrence of the clustered properties due to underlying causal mechanisms that accounts for the kindhood, but their functionally complementary causal interaction which produces unique features. One of the strengths of the synergic kinds account is its capacity to accommodate complex scientific categories that challenge traditional natural kind theories. Polymorphic species and now social institutions can be better understood under the synergic model, which allows for varied but functionally integrated properties to jointly constitute a kind’s identity. In the case of social institutions it can be so, for it accommodates both “indifferent” and “interactive” kindhoods with mechanisms of reinforcement learning, ecological feedback and social learning on the one side and the mechanism of belief-based population-level adaptation on the other side. Martínez also argues that the naturalness of a kind can be understood in terms of how its synergistic properties support reliable inductive inferences and scientific generalizations. This perspective supports a gradualist view of natural kinds, contrasting with rigid dichotomies between natural and non-natural (social) kinds, and aligns more closely with scientific practice.

The model also invites reflection through the lens of *ontic structural realism* (OSR) which says that fundamental ontology of science is not made up of individual objects, but of *relational structures* (Ladyman et al. 2007). In the context of social science, this suggests that what is most real about the social world is not the agents and their properties (even aggregate) but the *networks of roles, norms, and strategies* that structure their interaction.

The model embodies this structurally realist intuition:

* Agents succeed not by virtue of intrinsic properties, but by occupying *positions in a relational structure* like the roles of “owner” or “intruder,” inferred from cues and stabilized by mutually reinforcing strategies;
* The stability of these roles depends not on individual essence, but on *systemic regularities* characteristic of OSR.
* What emerges as real, and what constrains explanation, is not the psychology of any single agent but the *structured convention-space* of possible stable configurations, which is objective yet relative to concrete sets of normative beliefs which can be explained evolutionarliy.

In this way, the model bridges mechanistic explanation with structural realism: it shows how *relational regularities* encoded in role-structured strategies and cue mappings constitute the real explanatory grain of the social.

### Towards radically naturalistic metaphysics and social ontology

Ladyman et al. (2007) and Ross (2023) advance a *radically naturalistic metaphysics*, which holds that metaphysics must be continuous with empirical science and justified by its explanatory utility within mature scientific theories. They reject metaphysical speculation untethered from the practices of scientific modeling, proposing instead that *science itself determines what exists*.

Applying this to social ontology, we can make a stronger claim: if conventions and norms emerge in agent-based models with the same ontic status as genes in population genetics or strategies in evolutionary ecology, then *they qualify as part of the ontological inventory of a naturalized social science*.

This stands in sharp contrast to sociological paradigms that treat social structures as interpretive, symbolic, contingent or unbound, without requiring them to be *ontically robust or mechanistically instantiated*. Instead, I argue for:

* *Metaphysical minimalism and explanatory necessity*: if evolved (proto)-normative conventions explain the stability of social institutions and structure interaction, they belong in the ontology;
* *Empirical accountability of social ontology*: social kinds must be discoverable through modeling, experimentation, or observation, and their legitimacy depends on their integration into causal mechanistic models.

This enforces *ontic constraints on social ontology and sociological theory*, akin to those in biology or physics. Social kinds that cannot be grounded in such models due to lackling causal power, explanatory integration, or evolutionary plausibility might be regarded with skepticism or even discarded altogether.

By combining scientific realism, ontic explanation, structural realism, and radical naturalism, we arrive at a new vision of social scientific ontology:

* *Realist*: social structures are real if and only if they have causal roles in empirically adequate theories;
* *Mechanistic*: social explanations are credible when they trace outcomes to specific, instantiated processes—like learning, selection, and cue-driven inference;
* *Structural*: the ontology of the social is fundamentally about patterns, roles, and relationships and not individual actors as such, although it is scale-relative and exhibits “multi-level” real patterns;
* *Naturalistic*: Social kinds must arise from, and be constrained by, the same evolutionary and informational principles that structure the rest of nature.

This vision moves beyond interpretivism and beyond naive positivism. It offers a *scientifically mature metaphysics of the social*, anchored in computational models, empirical constraints, and evolutionary principles.

# References

Albert, Réka, and Juilee Thakar. 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): 353–69. <https://doi.org/10.1002/wsbm.1273>.

Alexander, J. McKenzie. 2009. “Social Deliberation: Nash, Bayes, and the Partial Vindication of Gabriele Tarde.” *Episteme* 6 (2): 164–84. <https://doi.org/10.3366/E1742360009000628>.

Alexander, Jason, and Brian Skyrms. 1999. “Bargaining with Neighbors: Is Justice Contagious?” *The Journal of Philosophy* 96 (11): 588. <https://doi.org/10.2307/2564625>.

Aoki, Masahiko. 2007. “Endogenizing Institutions and Institutional Changes\*.” *Journal of Institutional Economics* 3 (1): 1–31. <https://doi.org/10.1017/S1744137406000531>.

———. 2011. “Institutions as Cognitive Media Between Strategic Interactions and Individual Beliefs.” *Journal of Economic Behavior & Organization*, The Dynamics of Institutions in Perspectives: Alternative Conceptions and Future Challenges, 79 (1): 20–34. <https://doi.org/10.1016/j.jebo.2011.01.025>.

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

Aumann, Robert 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>.

———. 1976. “Agreeing to Disagree.” *The Annals of Statistics* 4 (6). <https://doi.org/10.1214/aos/1176343654>.

———. 1987. “Correlated Equilibrium as an Expression of Bayesian Rationality.” *Econometrica* 55 (1): 1. <https://doi.org/10.2307/1911154>.

Bacharach, M., and M. Bernasconi. 1997. “The Variable Frame Theory of Focal Points: An Experimental Study.” *Games and Economic Behavior* 19 (1): 1–45. <https://doi.org/10.1006/game.1997.0546>.

Bain, Robert. 2016. “Are Our Brains Bayesian?” *Significance* 13 (4): 14–19. <https://doi.org/10.1111/j.1740-9713.2016.00935.x>.

Baker, Lynne Rudder. 2019. “Just What Is Social Ontology?” *Journal of Social Ontology* 5 (1): 1–12. <https://doi.org/10.1515/jso-2019-2001>.

Baraghith, Karim. 2019. “Emergence of Public Meaning from a Teleosemantic and Game Theoretical Perspective.” *Journal of Philosophy*, 30.

Bardone, Emanuele, and Lorenzo Magnani. 2007. “Sharing Representations Through Cognitive Niche Construction.” *Data Science Journal* 6: S87–91. <https://doi.org/10.2481/dsj.6.S87>.

Bardsley, Nicholas, Judith Mehta, Chris Starmer, and Robert Sugden. 2010. “Explaining Focal Points: Cognitive Hierarchy Theory Versus Team Reasoning.” *The Economic Journal* 120 (543): 40–79. <https://doi.org/10.1111/j.1468-0297.2009.02304.x>.

Bechtel, William, and Adele Abrahamsen. 2005. “Explanation: A Mechanist Alternative.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, Mechanisms in biology, 36 (2): 421–41. <https://doi.org/10.1016/j.shpsc.2005.03.010>.

Bedau, Mark A. 1996. “The Extent to Which Organisms Construct Their Environments.” *Adaptive Behavior* 4 (3-4): 476–82.

Beni, Majid Davoody. 2017. “Structural Realism, Metaphysical Unification, and the Ontology and Epistemology of Patterns.” *International Studies in the Philosophy of Science* 31 (3): 285–300. <https://doi.org/10.1080/02698595.2018.1463691>.

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

Bicchieri, Cristina, and Peter McNally. 2018. “SHRIEKING SIRENS: SCHEMATA, SCRIPTS, AND SOCIAL NORMS. HOW CHANGE OCCURS.” *Social Philosophy and Policy* 35 (1): 23–53. <https://doi.org/10.1017/S0265052518000079>.

Bicchieri, Cristina, Ryan Muldoon, and Alessandro Sontuoso. 2023. “Social Norms.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Winter 2023. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2023/entries/social-norms/>.

Bickhard, Mark H. 2008. “Social Ontology as Convention.” *Topoi* 27 (1-2): 139–49. <https://doi.org/10.1007/s11245-008-9036-1>.

Binmore, Ken. 2008. “Do Conventions Need to Be Common Knowledge?” *Topoi* 27 (1): 17. <https://doi.org/10.1007/s11245-008-9033-4>.

Bishop, D.T., and C. Cannings. 1978. “A Generalized War of Attrition.” *Journal of Theoretical Biology* 70 (1): 85–124. <https://doi.org/10.1016/0022-5193(78)90304-1>.

Bonnay, Denis, and Paul Egré. 2009. “Inexact Knowledge with Introspection.” *Journal of Philosophical Logic* 38: 179–227.

Bowles, Samuel, Jung-Kyoo Choi, and Astrid Hopfensitz. 2003. “The Co-Evolution of Individual Behaviors and Social Institutions.” *Journal of Theoretical Biology* 223 (2): 135–47. <https://doi.org/10.1016/S0022-5193(03)00060-2>.

Boyd, Richard. 1991. “Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds.” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 61 (1/2): 127–48. <https://www.jstor.org/stable/4320174>.

———. 2021. “Rethinking Natural Kinds, Reference and Truth: Towards More Correspondence with Reality, Not Less.” *Synthese* 198 (12): 2863–903. <https://doi.org/10.1007/s11229-019-02138-4>.

Boyd, Richard N. 1983. “On the Current Status of the Issue of Scientific Realism.” In *Methodology, Epistemology, and Philosophy of Science: Essays in Honour of Wolfgang Stegmüller on the Occasion of His 60th Birthday, June 3rd, 1983*, edited by Carl G. Hempel, Hilary Putnam, and Wilhelm K. Essler, 45–90. Dordrecht: Springer Netherlands. <https://doi.org/10.1007/978-94-015-7676-5_3>.

Bratman, Michael. 2022. *Shared and Institutional Agency: Toward a Planning Theory of Human Practical Organization*. New York, NY, United States of America: Oxford University Press.

Burge, Tyler. 1975. “On Knowledge and Convention.” *The Philosophical Review* 84 (2): 249. <https://doi.org/10.2307/2183970>.

Chakravartty, Anjan. 2007. *A metaphysics for scientific realism: knowing the unobservable*. 1. publ. Cambridge: Cambridge University Press.

Churchland, Paul M. 1992. *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. MIT Press.

Craver, Carl F. 2014. “The Ontic Account of Scientific Explanation.” In *Explanation in the Special Sciences: The Case of Biology and History*, edited by Marie I. Kaiser, Oliver R. Scholz, Daniel Plenge, and Andreas Hüttemann, 27–52. Synthese Library. Dordrecht: Springer Netherlands. <https://doi.org/10.1007/978-94-007-7563-3_2>.

Crawford, Sue E. S., and Elinor Ostrom. 1995. “A Grammar of Institutions.” *American Political Science Review* 89 (3): 582–600. <https://doi.org/10.2307/2082975>.

Cripps, Martin. 1991. “Correlated Equilibria and Evolutionary Stability.” *Journal of Economic Theory* 55 (2): 428–34. <https://doi.org/10.1016/0022-0531(91)90048-9>.

Cubitt, Robin P., and Robert Sugden. 2003. “COMMON KNOWLEDGE, SALIENCE AND CONVENTION: A RECONSTRUCTION OF DAVID LEWIS’ GAME THEORY.” *Economics & Philosophy* 19 (2): 175–210. <https://doi.org/10.1017/S0266267103001123>.

Davies, N.B. 1978. “Territorial Defence in the Speckled Wood Butterfly (Pararge Aegeria): The Resident Always Wins.” *Animal Behaviour* 26 (February): 138–47. <https://doi.org/10.1016/0003-3472(78)90013-1>.

Dennett, Daniel C. 1991. “Real Patterns.” *The Journal of Philosophy* 88 (1): 27–51.

Elder-Vass, Dave. 2007. “A Method for Social Ontology: Iterating Ontology and Social Research.” *Journal of Critical Realism* 6 (2): 226–49. <https://doi.org/10.1558/jocr.v6i2.226>.

———. 2010. *The Causal Power of Social Structures: Emergence, Structure and Agency*. Cambridge University Press.

Elster, Jon. 1989. *The Cement of Society: A Survey of Social Order*. Cambridge University Press. <https://philpapers.org/rec/ELSTCO-2>.

Engel, Christoph, and W. Singer, eds. 2008. *Better Than Conscious? Decision Making, the Human Mind, and Implications for Institutions*. Strüngmann Forum Reports. Cambridge, Mass: MIT Press.

Epstein, Brian. 2015. *The Ant Trap: Rebuilding the Foundations of the Social Sciences*. Oxford Studies in Philosophy of Science. New York, NY: Oxford University Press.

———. 2016. “A Framework for Social Ontology.” *Philosophy of the Social Sciences* 46 (2): 147–67. <https://doi.org/10.1177/0048393115613494>.

———. 2018. “Social Ontology.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2018. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/sum2018/entries/social-ontology/>.

Epstein, Joshua M, and Robert Axtell. 1996. *Growing Artificial Societies: Social Science from the Bottom up*. Brookings Institution Press.

Erev, Ido, and Alvin E. Roth. 1998. “Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria.” *The American Economic Review* 88 (4): 848–81. <https://www.jstor.org/stable/117009>.

Ferguson, Adam. (1767) 1980. *An Essay on the History of Civil Society, 1767*. Transaction Publishers.

Franke, Michael, and Elliott O. Wagner. 2014. “Game Theory and the Evolution of Meaning: Game Theory and the Evolution of Meaning.” *Language and Linguistics Compass* 8 (9): 359–72. <https://doi.org/10.1111/lnc3.12086>.

Friston, Karl. 2010. “The Free-Energy Principle: A Unified Brain Theory?” *Nature Reviews Neuroscience* 11 (2): 127–38. <https://doi.org/10.1038/nrn2787>.

Friston, Karl, Philipp Schwartenbeck, Thomas Fitzgerald, Michael Moutoussis, Tim Behrens, and Raymond Dolan. 2013. “The Anatomy of Choice: Active Inference and Agency.” *Frontiers in Human Neuroscience* 7. <https://www.frontiersin.org/articles/10.3389/fnhum.2013.00598>.

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

Gigerenzer, Gerd, and Reinhard Selten. 2002. *Bounded Rationality: The Adaptive Toolbox*. MIT Press.

Gigerenzer, Gerd, and Peter M. Todd. 1999. *Simple Heuristics That Make Us Smart*. Evolution and Cognition. New York: Oxford University Press.

Gilbert, Margaret. 1992. *On Social Facts*. Princeton University Press.

Gintis, Herbert. 2007a. “A Framework for the Unification of the Behavioral Sciences.” *Behavioral and Brain Sciences* 30 (1): 1–16. <https://doi.org/10.1017/S0140525X07000581>.

———. 2007b. “The Evolution of Private Property.” *Journal of Economic Behavior & Organization* 64 (1): 1–16. <https://doi.org/10.1016/j.jebo.2006.02.002>.

———. 2009a. *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Interaction*. 2nd ed. Princeton: Princeton University Press.

———. 2009b. *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton, N.J: Princeton University Press.

Gintis, Herbert, and Dirk Helbing. 2013. “Homo Socialis: An Analytical Core for Sociological Theory.” *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2362262>.

Godfrey-Smith, Peter. 1991. “Signal, Decision, Action.” *The Journal of Philosophy* 88 (12): 709–22. <https://doi.org/10.2307/2027008>.

———. 2002. “On the Evolution of Representational and Interpretive Capacities.” *The Monist* 85 (1): 50–69. <https://www.jstor.org/stable/27903757>.

———. 2003. *Theory and Reality: An Introduction to the Philosophy of Science*. Science and Its Conceptual Foundations. Chicago: University of Chicago Press.

———. 2014. “Sender-Receiver Systems Within and Between Organisms.” *Philosophy of Science* 81 (5): 866–78. <https://doi.org/10.1086/677686>.

———. 2020. “In the Beginning There Was Information?” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 80 (April): 101239. <https://doi.org/10.1016/j.shpsc.2019.101239>.

———. n.d. “Complexity and the Function of Mind in Nature (Cambridge Studies in Philosophy and Biology),” 329.

Godfrey-Smith, Peter, In R Sternberg, and J Kaufman. n.d. “Environmental Complexity and the Evolution of Cognition,” 27.

Gold, Natalie, and Robert Sugden. 2007. “Collective Intentions and Team Agency.” *The Journal of Philosophy* 104 (3): 109–37. <https://www.jstor.org/stable/20620005>.

Guala, Francesco. 2007. “The Philosophy of Social Science: Metaphysical and Empirical.” *Philosophy Compass* 2 (6): 954–80. <https://doi.org/10.1111/j.1747-9991.2007.00095.x>.

———. 2016. *Understanding Institutions: The Science and Philosophy of Living Together*. Princeton: Princeton University Press.

———. 2019. “Social Norms, Expectations and Sanctions.” *Analyse & Kritik* 41 (2): 375–82. <https://doi.org/10.1515/auk-2019-0023>.

———. 2020. “Solving the Hi-Lo Paradox: Equilibria, Beliefs, and Coordination.” In *Minimal Cooperation and Shared Agency*, edited by Anika Fiebich, 11:149–68. Cham: Springer International Publishing. <https://doi.org/10.1007/978-3-030-29783-1_9>.

Guala, Francesco, and Frank Hindriks. 2015. “A UNIFIED SOCIAL ONTOLOGY.” *The Philosophical Quarterly* 65 (259): 177–201. <https://doi.org/10.1093/pq/pqu072>.

Guala, Francesco, and Luigi Mittone. 2010. “How History and Convention Create Norms: An Experimental Study.” *Journal of Economic Psychology* 31 (4): 749–56. <https://doi.org/10.1016/j.joep.2010.05.009>.

Hacking, Ian. 1999. *The Social Construction of What?* Harvard University Press.

Harms, William F. 2004. *Information and Meaning in Evolutionary Processes*. Cambridge University Press.

Haslanger, Sally Anne. 2012. *Resisting Reality: Social Construction and Social Critique*. New York: Oxford University Press.

Hédoin, Cyril. 2016. “Toward a Naturalistic Philosophy of Institutions.” *Œconomia. History, Methodology, Philosophy*, no. 6-3 (September): 443–50. <https://journals.openedition.org/oeconomia/2400>.

———. 2021. “The Beliefs-Rules-Equilibrium Account of Institutions: A Contribution to a Naturalistic Social Ontology.” *Journal of Social Ontology* 7 (1): 73–96. <https://doi.org/10.1515/jso-2020-0001>.

Hedström, Peter, and Peter S. Bearman, eds. 2009. *The Oxford Handbook of Analytical Sociology*. Oxford Handbooks. Oxford ; New York: Oxford University Press.

Herrmann, Daniel Alexander, and Brian Skyrms. 2021. “Invention and Evolution of Correlated Conventions.” *The British Journal for the Philosophy of Science*, September. <https://doi.org/10.1086/717161>.

Herrmann, Daniel Alexander, and Jacob VanDrunen. 2022. “Sifting the Signal from the Noise.” *The British Journal for the Philosophy of Science*, May, 720805. <https://doi.org/10.1086/720805>.

Hindriks, F. 2019. “Norms That Make a Difference: Social Practices and Institutions.” *Analyse Und Kritik* 41 (1): 125–45. <https://doi.org/10.1515/auk-2019-410109>.

Hindriks, Frank. 2005. *Rules & Institutions: Essays on Meaning, Speech Acts and Social Ontology: Essays over Betekenis, Taalhandelingen En Sociale Ontologie = Regels & Instituties*. Alblasserdam: Haveka BV.

———. 2022. “Institutions and Their Strength.” *Economics & Philosophy* 38 (3): 354–71. <https://doi.org/10.1017/S0266267121000195>.

Hindriks, Frank, and Francesco Guala. 2015. “Institutions, Rules, and Equilibria: A Unified Theory\*.” *Journal of Institutional Economics* 11 (3): 459–80. <https://doi.org/10.1017/S1744137414000496>.

———. 2021. “The Functions of Institutions: Etiology and Teleology.” *Synthese* 198 (3): 2027–43. <https://doi.org/10.1007/s11229-019-02188-8>.

Hobbes, Thomas. (1650) 2016. *Thomas Hobbes: Leviathan (Longman Library of Primary Sources in Philosophy)*. Edited by Marshall Missner. 0th ed. Routledge. <https://doi.org/10.4324/9781315507613>.

Hodder, Ian. 2014. “The Entanglements of Humans and Things: A Long-Term View.” *New Literary History* 45 (1): 19–36. <https://doi.org/10.1353/nlh.2014.0005>.

Hume, David. 1998. *An Enquiry Concerning the Principles of Morals: A Critical Edition*. Vol. 4. Oxford University Press.

———. 2003. *A Treatise of Human Nature*. Courier Corporation.

Huttegger, Simon M. 2007a. “Evolution and the Explanation of Meaning.” *Philosophy of Science* 74 (1): 1–27. <https://doi.org/10.1086/519477>.

———. 2007b. “Evolutionary Explanations of Indicatives and Imperatives.” *Erkenntnis* 66 (3): 409–36. <https://doi.org/10.1007/s10670-006-9022-1>.

Huttegger, Simon M., and Brian Skyrms. 2008. “Emergence of Information Transfer by Inductive Learning.” *Studia Logica* 89 (2): 237–56. <https://doi.org/10.1007/s11225-008-9123-8>.

Izquierdo, Segismundo S., Luis R. Izquierdo, and Christoph Hauert. 2024. “Positive and Negative Selective Assortment in Pairwise Interactions.” bioRxiv. <https://doi.org/10.1101/2024.07.18.604169>.

Jankovic, Marija, and Kirk Ludwig, eds. 2018. *The Routledge handbook of collective intentionality*. Routledge handbooks in philosophy. London New York: Routledge, Taylor & Francis Group. <https://doi.org/10.4324/9781315768571>.

Jensen, Martin Kaae, and Alexandros Rigos. 2018. “Evolutionary Games and Matching Rules.” *International Journal of Game Theory* 47 (3): 707–35. <https://doi.org/10.1007/s00182-018-0630-1>.

Kincaid, Harold. 2024. “A Naturalist Approach to Social Ontology.” *Synthese* 203 (1): 27. <https://doi.org/10.1007/s11229-023-04432-8>.

Kitcher, Philip. 1981. “Explanatory Unification.” *Philosophy of Science* 48 (4): 507–31. <https://doi.org/10.1086/289019>.

———. n.d. “Explanatory Unification and the Causal Structure of the World.”

Kôno, Norio. 2008. “Noncooperative Game in Cooperation: Reformulation of Correlated Equilibria1).”

Korbak, Tomasz, Julian Zubek, Łukasz Kuciński, Piotr Miłoś, and Joanna Rączaszek-Leonardi. 2021. “Interaction History as a Source of Compositionality in Emergent Communication.” *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems* 22 (2): 212–43. <https://doi.org/10.1075/is.21020.kor>.

LaCroix, Travis. 2020a. “Evolutionary Explanations of Simple Communication: Signalling Games and Their Models.” *Journal for General Philosophy of Science* 51 (1): 19–43. <https://doi.org/10.1007/s10838-019-09481-7>.

———. 2020b. “On Salience and Signaling in Sender–Receiver Games: Partial Pooling, Learning, and Focal Points.” *Synthese* 197 (4): 1725–47. <https://doi.org/10.1007/s11229-018-1766-z>.

Ladyman, James, Don Ross, David Spurrett, and John Collier. 2007. *Every Thing Must Go: Metaphysics Naturalized*. OUP Oxford.

Lauer, Richard. 2019. “Is Social Ontology Prior to Social Scientific Methodology?” *Philosophy of the Social Sciences* 49 (3): 171–89. <https://doi.org/10.1177/0048393119840328>.

Lewis, David. (1969) 2008. *Convention: A Philosophical Study*. John Wiley & Sons.

Lismont, Luc, and Philippe Mongin. 1995. “Belief Closure: A Semantics of Common Knowledge for Modal Propositional Logic.” *Mathematical Social Sciences* 30 (2): 127–53.

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

Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. “Thinking about Mechanisms.” *Philosophy of Science* 67 (1): 1–25. <https://doi.org/10.1086/392759>.

Marmor, Andrei. 1996. “On Convention.” *Synthese* 107 (3): 349–71. <https://doi.org/10.1007/BF00413841>.

———. 2009. *Social Conventions: From Language to Law*. Princeton Monographs in Philosophy. Princeton, N.J: Princeton University Press.

Martínez, Manolo. 2019. “Deception as Cooperation.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 77 (October): 101184. <https://doi.org/10.1016/j.shpsc.2019.101184>.

———. 2020. “Synergic Kinds.” *Synthese* 197 (5): 1931–46. <https://doi.org/10.1007/s11229-017-1480-2>.

Martínez, Manolo, and Peter Godfrey-Smith. 2016. “Common Interest and Signaling Games: A Dynamic Analysis.” *Philosophy of Science* 83 (3): 371–92. <https://doi.org/10.1086/685743>.

Maynard Smith, John. 1982. *Evolution and the Theory of Games*. Cambridge ; New York: Cambridge University Press.

Metzger, Lars P. 2018. “Evolution and Correlated Equilibrium.” *Journal of Evolutionary Economics* 28 (2): 333–46. <https://doi.org/10.1007/s00191-017-0539-z>.

Millikan, Ruth G. 2022. “Language Conventions Made Simple.” *The Journal of Philosophy*, 21.

Millikan, Ruth Garrett. 1987. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press.

———. 2005. *Language: A Biological Model*. Oxford : Oxford ; New York: Clarendon Press ; Oxford University Press.

Millikan, Ruth Garrett, and Ruth G. Millikan. 2004. *Varieties of Meaning: The 2002 Jean Nicod Lectures*. MIT Press.

Mithen, Steven J. 1988. “Looking and Learning: Upper Palaeolithic Art and Information Gathering.” *World Archaeology* 19 (3): 297–327. <https://doi.org/10.1080/00438243.1988.9980043>.

Monderer, Dov, and Dov Samet. 1989. “Approximating Common Knowledge with Common Beliefs.” *Games and Economic Behavior* 1 (2): 170–90.

Morris, Stephen. 1999. “Approximate Common Knowledge Revisited.” *International Journal of Game Theory* 28: 385–408.

Morsky, Bryce, and Erol Akçay. 2019. “Evolution of Social Norms and Correlated Equilibria.” *Proceedings of the National Academy of Sciences* 116 (18): 8834–39. <https://doi.org/10.1073/pnas.1817095116>.

Mouden, Claire El, Maxwell Burton-Chellew, Andy Gardner, and Stuart A. West. 2012. “What Do Humans Maximize?” In *Evolution and Rationality: Decisions, Co-Operation and Strategic Behaviour*, edited by Ken Binmore and Samir Okasha, 23–49. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511792601.003>.

Mussweiler, Thomas, Andrew R Todd, and Jan Crusius. 2012. “The Cognitive Underpinnings of Social Behavior.” In *Evolution and the Mechanisms of Decision Making*, 12. MIT Press.

North, Douglas. 1990. *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press.

O’Connor, Cailin. 2019. *The Origins of Unfairness: Social Categories and Cultural Evolution*. First edition. Oxford ; New York, NY: Oxford University Press.

———. 2020. *Games in the Philosophy of Biology*. 1st ed. Cambridge University Press. <https://doi.org/10.1017/9781108616737>.

———. 2021. “Measuring Conventionality.” *Australasian Journal of Philosophy* 99 (3): 579–96. <https://doi.org/10.1080/00048402.2020.1781220>.

Odling-Smee, F. John, Kevin N. Lala, and Marcus Feldman. 2003. “Niche Construction: The Neglected Process in Evolution.” In *Niche Construction*. Princeton University Press. <https://www.degruyterbrill.com/document/doi/10.1515/9781400847266/html>.

Okasha, S. 2017. “Biology and the Theory of Rationality.” In. <https://doi.org/10.1017/9781107295490.009>.

Okasha, Samir. 2015. “The Evolution of Bayesian Updating.” *Philosophy of Science*, July. <https://doi.org/10.1086/674058>.

Okasha, Samir, and Ken Binmore. 2012. *Evolution and Rationality: Decisions, Co-Operation and Strategic Behaviour*. Cambridge University Press.

Parsons, Talcott. 2015. “The Place of Ultimate Values in Sociological Theory.” *The International Journal of Ethics*, September. <https://doi.org/10.1086/intejethi.45.3.2378271>.

Paternotte, Cedric. 2011. “Being Realistic about Common Knowledge: A Lewisian Approach.” *Synthese* 183 (2): 249–76. <https://doi.org/10.1007/s11229-010-9770-y>.

Paternotte, Cédric. 2020. “Social Evolution and the Individual-as-Maximising-Agent Analogy.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 79 (February): 101225. <https://doi.org/10.1016/j.shpsc.2019.101225>.

Perea, Andrés. n.d. “Epistemic Game Theory,” 582.

Planer, Ronald J., and Peter Godfrey‐Smith. 2021. “Communication and Representation Understood as Sender–Receiver Coordination.” *Mind & Language* 36 (5): 750–70. <https://doi.org/10.1111/mila.12293>.

Psillos, Stathis. 1999. *Scientific Realism: How Science Tracks Truth*. 0th ed. Routledge. <https://doi.org/10.4324/9780203979648>.

Pufendorf, Samuel. 1673. *De Officio Hominis et civis juxta legem naturalem libri duo*. Lund: Junghans. <https://catalogue-bu.u-bourgogne.fr/discovery/fulldisplay/alma991001697579706659/33UB_INST:33UB_INST>.

Rabinowicz, Wlodek. 2018. “Are Institutions Rules in Equilibrium? Comments on Guala’s *Understanding Institutions*.” *Philosophy of the Social Sciences* 48 (6): 569–84. <https://doi.org/10.1177/0048393118798643>.

Rakoczy, Hannes, and Michael Tomasello. 2007. “The Ontogeny of Social Ontology: Steps to Shared Intentionality and Status Functions.” In *Intentional Acts and Institutional Facts: Essays on John Searle’s Social Ontology*, edited by Savas L. Tsohatzidis, 113–37. Theory and Decision Library. Dordrecht: Springer Netherlands. <https://doi.org/10.1007/978-1-4020-6104-2_5>.

Rescorla, Michael. 2024. “Convention.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Spring 2024. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2024/entries/convention/>.

Ross, Don. 2012. “Coordination and the Foundations of Social Intelligence.” The Oxford Handbook of Philosophy of Social Science. August 6, 2012. <https://doi.org/10.1093/oxfordhb/9780195392753.013.0020>.

———. 2023. “Scientific Metaphysics and Social Science.” *Synthese* 202 (October). <https://doi.org/10.1007/s11229-023-04358-1>.

Samir Okasha, and Samir Okasha. 2013. “The Evolution of Bayesian Updating.” *Philosophy of Science* 80 (5): 745–57. <https://doi.org/10.1086/674058>.

Sarkia, Matti, and Tuukka Kaidesoja. 2023. “Two Approaches to Naturalistic Social Ontology.” *Synthese* 201 (3): 104. <https://doi.org/10.1007/s11229-023-04105-6>.

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

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

Schiffer, Stephen R. 1972. *Meaning*. Oxford,: Clarendon Press.

Schliesser, Eric. 2024. “Locke’s Humean Conventionalism.” *British Journal for the History of Philosophy* 32 (6): 1457–66. <https://doi.org/10.1080/09608788.2023.2268404>.

Schmid, Hans Bernhard. 2023. *We, Together: The Social Ontology of Us*. 1st ed. Oxford University PressNew York. <https://doi.org/10.1093/oso/9780197563724.001.0001>.

Searle, John. 1995. *The Construction of Social Reality*. Simon and Schuster.

———. 2010. *Making the Social World: The Structure of Human Civilization*. Oxford University Press.

Searle, John R. 2015. “Status Functions and Institutional Facts: Reply to Hindriks and Guala.” *Journal of Institutional Economics* 11 (3): 507–14. <https://doi.org/10.1017/S1744137414000629>.

Seitz, Fabian. 2020. “Argumentation Evolved: But How? Coevolution of Coordinated Group Behavior and Reasoning.” *Argumentation* 34 (2): 237–60. <https://doi.org/10.1007/s10503-020-09510-6>.

Seth, Anil K. 2002. “Agent-Based Modelling and the Environmental Complexity Thesis.” In *From Animals to Animats 7*, edited by Bridget Hallam, Dario Floreano, John Hallam, Jean-Arcady Meyer, and Gillian Hayes, 13–24. The MIT Press. <https://doi.org/10.7551/mitpress/3121.003.0004>.

Seyfarth, Robert, and Dorothy Cheney. 1990. “The Assessment by Vervet Monkeys of Their Own and Another Species’ Alarm Calls.” *Animal Behaviour* 40 (4): 754–64. <https://doi.org/10.1016/S0003-3472(05)80704-3>.

Shannon, Claude E. 1948. “A Mathematical Theory of Communication.” *The Bell System Technical Journal* 27 (3): 379–423.

Shea, Nicholas, Peter Godfrey-Smith, and Rosa Cao. 2018. “Content in Simple Signalling Systems.” *The British Journal for the Philosophy of Science* 69 (4): 1009–35. <https://doi.org/10.1093/bjps/axw036>.

Simons, Mandy, and Kevin J. S. Zollman. 2019. “Natural Conventions and Indirect Speech Acts.” *Philosopher’s Imprint* 19 (9). <http://hdl.handle.net/2027/spo.3521354.0019.009>.

Skyrms, Brian. 1994. “Darwin Meets the Logic of Decision: Correlation in Evolutionary Game Theory.” *Philosophy of Science* 61 (4): 503–28. <https://doi.org/10.1086/289819>.

———. 2003. *The Stag Hunt and the Evolution of Social Structure*. 1st ed. Cambridge University Press. <https://doi.org/10.1017/CBO9781139165228>.

———. 2010a. *Signals: Evolution, Learning, & Information*. Oxford ; New York: Oxford University Press.

———. 2010b. “The Flow of Information in Signaling Games.” *Philosophical Studies* 147 (1): 155–65. <https://doi.org/10.1007/s11098-009-9452-0>.

———. 2012. “Learning to Signal with Probe and Adjust.” *Episteme* 9 (2): 139–50. <https://doi.org/10.1017/epi.2012.5>.

———. 2014a. *Evolution of the Social Contract*. Second edition. Cambridge ; New York: Cambridge University Press.

———. 2014b. *Social Dynamics*. Oxford: Oxford University Press.

———. 2023. “Quasi-Conventions.” *Synthese* 201 (3): 99. <https://doi.org/10.1007/s11229-023-04103-8>.

Skyrms, Brian, and Robin Pemantle. 2004. “A Dynamic Model of Social Network Formation.” arXiv. <https://doi.org/10.48550/arXiv.math/0404101>.

Smith, J. Maynard, and G. R. Price. 1973. “The Logic of Animal Conflict.” *Nature* 246 (5427): 15–18. <https://doi.org/10.1038/246015a0>.

Smith, John Maynard, and G. A. Parker. 1976. “The Logic of Asymmetric Contests.” *Animal Behaviour* 24 (1): 159–75. <https://doi.org/10.1016/S0003-3472(76)80110-8>.

Solan, Eilon, and Rakesh V Vohra. n.d. “Correlated Equilibrium, Public Signalling and Absorbing Games.”

Sperber, Dan. 1996. *Explaining Culture: A Naturalistic Approach*. Oxford, UK ; Cambridge, Mass: Blackwell.

Sperber, Dan, and Deirdre Wilson. 2002. “Pragmatics, Modularity and Mind-Reading.” *Mind & Language* 17 (1-2): 3–23. https://doi.org/<https://doi.org/10.1111/1468-0017.00186>.

Sperry-Taylor, Ashton T. 2021. “Reassessing Equilibrium Explanations: When Are They Causal Explanations?” *Synthese* 198 (6): 5577–98. <https://doi.org/10.1007/s11229-019-02423-2>.

Sterelny, Kim. 2003. *Thought in a Hostile World: The Evolution of Human Cognition*. Malden, MA: Blackwell.

———. 2012a. “From Fitness to Utility.” In *Evolution and Rationality*, edited by Ken Binmore and Samir Okasha, 246–73. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511792601.012>.

———. 2012b. *The Evolved Apprentice: How Evolution Made Humans Unique*. Jean Nicod Lectures. Cambridge, Mass: The MIT Press.

———. 2012c. “A Glass Half-Full: Brian Skyrms’s ‘Signals’.” *Economics and Philosophy* 28 (1): 73–86. <https://doi.org/10.1017/S0266267112000120>.

———. 2016. “Cooperation, Culture, and Conflict.” *The British Journal for the Philosophy of Science* 67 (1): 31–58. <https://doi.org/10.1093/bjps/axu024>.

———. 2017. “From Code to Speaker Meaning.” *Biology & Philosophy* 32 (6): 819–38. <https://doi.org/10.1007/s10539-017-9597-8>.

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

Sugden, Robert. 2005. *The Economics of Rights, Co-Operation and Welfare*. London: Palgrave Macmillan UK. <https://doi.org/10.1057/9780230536791>.

Sugden, Robert, and Ignacio E. Zamarrón. 2006. “Finding the Key: The Riddle of Focal Points.” *Journal of Economic Psychology* 27 (5): 609–21. <https://doi.org/10.1016/j.joep.2006.04.003>.

Tuomela, Raimo. 2002. *The Philosophy of Social Practices: A Collective Acceptance View*. 1st ed. Cambridge University Press. <https://doi.org/10.1017/CBO9780511487446>.

———. 2013. *Social Ontology: Collective Intentionality and Group Agents*. Oxford University Press.

Ullmann-Margalit, Edna. 1977. *The Emergence of Norms*. Clarendon Library of Logic and Philosophy. Oxford [Eng]: Clarendon Press.

Vanderschraaf, P. 1998. “Knowledge, Equilibrium and Convention.” *Erkenntnis* 49 (3): 337–69. <https://doi.org/10.1023/A:1005461514200>.

Vanderschraaf, Peter. 1995. “Convention as Correlated Equilibrium.” *Erkenntnis* 42 (1): 65–87. <https://doi.org/10.1007/BF01666812>.

———. 2001. *Learning and Coordination: Inductive Deliberation, Equilibrium, and Convention*. Studies in Ethics. New York: Routledge.

———. 2018. “Learning Bargaining Conventions.” *Social Philosophy and Policy* 35 (1): 237–63. <https://doi.org/10.1017/S0265052518000110>.

Vanderschraaf, Peter, and Giacomo Sillari. 2023. “Common Knowledge.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Winter 2023. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2023/entries/common-knowledge/>.

Vanderschraaf, Peter, and Brian Skyrms. 1993. “Deliberational Correlated Equilibria.” *Philosophical Topics* 21 (1): 191–227. <https://www.jstor.org/stable/43154147>.

———. n.d. “Learning to Take Turns,” 38.

VanDrunen, Jacob, and Daniel A. Herrmann. 2023. “Naturalizing Natural Salience.” *British Journal for the Philosophy of Science*. <https://www.semanticscholar.org/paper/Naturalizing-Natural-Salience-VanDrunen-Herrmann/088559d154d4d106095c93324aaad8fd4f5ab62d>.

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

———. 2019. “The Evolution of Social Contracts.” *Journal of Social Ontology* 5 (2): 181–203. <https://doi.org/10.1515/jso-2019-0041>.

Weber, Max. 1924. *Gesammelte Aufsätze zur Soziologie und Sozialpolitik*. Tübingen: Mohr.

Wilson, Deirdre, and Dan Sperber. 1999. “Relevance and Relevance Theory.” In *MIT Encyclopedia of the Cognitive Sciences*, 719–22. MIT Press.

Woodward, James. 2005. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, USA.

Wright, Larry. 1973. “Functions.” *The Philosophical Review* 82 (2): 139. <https://doi.org/10.2307/2183766>.

Young, H. Peyton. 1998. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton, N.J: Princeton University Press.

Zachník, Vojtěch. 2021. “Epistemic Foundations of Salience-Based Coordination.” *Organon F* 2021 (4): 819–44. <https://doi.org/10.31577/orgf.2021.28404>.

Zamir, Shmuel, Michael Maschler, and Eilon Solan. 2013. *Game theory*. Cambridge: Cambridge University Press.

Zawidzki, Tadeusz Wiesław. 2013. *Mindshaping: A New Framework for Understanding Human Social Cognition*. The MIT Press. <https://doi.org/10.7551/mitpress/8441.001.0001>.

1. These are “homeostatic property clusters” (R. Boyd 1991), co-occurring properties that define kindhood. They support inductive generalization and scientific inference. [↑](#footnote-ref-20)
2. A solution concept from game theory generalizing Nash equilibrium, when all players get a private signal from a third party and it is in their best interest to follow that signal. A notable example is traffic lights. Guala and many other scholars claim social norms to be such correlation devices akin to traffic lights. [↑](#footnote-ref-21)
3. As Hume (1998) writes, “It has been asserted by some, that justice arises from human conventions, and proceeds from the voluntary choice, consent, or combination of mankind … if by convention be meant a sense of common interest; which sense each man feels in his own breast, which he remarks in his fellows, and which carries him, in concurrence with others, into a general plan or system of actions, which tends to public utility; it must be owned, that, in this sense, justice arises from human conventions. For if it be allowed (what is, indeed, evident) that the particular consequences of a particular act of justice may be hurtful to the public as well as to individuals; it follows, that every man, in embracing that virtue, must have an eye to the whole plan or system, and must expect the concurrence of his fellows in the same conduct and behaviour. Did all his views terminate in the consequences of each act of his own, his benevolence and humanity, as well as his self-love, might often prescribe to him measures of conduct very different from those, which are agreeable to the strict rules of right and justice …”. Schliesser (2024) notes that positive social externality is a requirement for a purely “Humean” convention. [↑](#footnote-ref-33)
4. A payoff matrix is a mathematical representation that shows the possible outcomes for each combination of strategies chosen by the players. Achieving coordination often requires stabilizing communication to arrive at mutual agreement, especially when different individuals or groups have conflicting preferences. This need for a reliable mechanism to resolve coordination issues is crucial in many social contexts. [↑](#footnote-ref-34)
5. Epistemic game theorists contend that there is no correlate of mixed-strategy equilibrium when viewed from epistemic (or knowledge) point of view (Perea, n.d.). I agree with them and only talk about it here for the purposes of comparison with CE. [↑](#footnote-ref-35)
6. On a planet identical to Earth in almost all respects but featuring water composed of XYZ rather than H₂O, inhabitants use the term “water” yet refer to different substance. According to Putnam, this illustrates that psychological states alone do not determine meaning; external factors like chemical composition and environmental acquisition influence linguistic reference. His assertion is encapsulated by his famous statement: “meanings just ain’t in the head”. This will be crucially important for us later in the discussion of the problem of ontic reference within the study of evolution of social conventions. [↑](#footnote-ref-37)
7. The emergence of objective yet relative moral norms in accordance with Lewisian approach and rigor was developed by (**mackenzie2007?**), which echoes “arbitrary yet stable” notion of instrumental conventions. [↑](#footnote-ref-38)
8. As Skyrms has shown (2010a, 2010b), the pattern can be learned dynamically in iterated games: both X and Y can be established and recognized with trial-and-error via reinforcement learning. [↑](#footnote-ref-40)
9. 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 and Thakar 2014), which is out of scope of this thesis. [↑](#footnote-ref-43)
10. A quite important clarification here is that to be “on the same page” about the need of common knowledge and the degree of rationality, we need to take into account the stage of a convention in question: is it just forming or is it already stable? It seems intuitive to suggest that earlier rounds of play require more explicit beliefs and cognitive demands than later rounds when strategies become more automatic and probabilities of actions of others are easier to predict. It is less costly to converge on an equilibrium in later rounds of play, so we need to be explicit about the state of play when discussing the need for common knowledge and cognitive demands of conventions. [↑](#footnote-ref-45)
11. Logicians Lismont and Mongin (1995) contend that epistemic interactions (or belief recursions) are not actual, but potential: agents do not compute them explicitly, but *could* deduce them from the situation give adequate cognitive resources. [↑](#footnote-ref-46)
12. However, seen dynamically, it can be argued the any convention came into being to solve a coordination problem, but after it have been established, it might have lost its initial coordinating function. [↑](#footnote-ref-48)
13. The Dirichlet rule is a Bayesian updating procedure based on the Dirichlet distribution used for modeling probabilities over a finite set of discrete outcomes (“a distribution over distributions”). In learning models, the Dirichlet rule updates the probability assigned to each probability distribution by counting the number of times each of them has produced a particular outcome such as a reward. These counts serve as parameters of the Dirichlet distribution, which then yields a probability distribution over the options. Formally, if option has been rewarded times, the updated probability for option is proportional to , and the probability vector over options is such that and . This rule captures how empirical frequencies shape probabilistic beliefs in a principled Bayesian manner. [↑](#footnote-ref-51)
14. The distinction between “exogenous” and “endogenous” information influencing agent’s strategy choice is already in Aumann (1987). The former type of information is obtained from external cues and the latter from agents’ reasoning about about how other agents reason. Aumann did not consider the distinction important, for the knowledge of exogeneity/endogeneity of agents’ information or even actions does not contribute to achieving CE. Vanderchraaf’s usage of Dirichlet dynamics clarified how endogeneity can contribute but did not eliminate the external signal altogether. [↑](#footnote-ref-52)
15. Many scholars use metaphors emphasizing the external character of CE: “mediator” and “correlation device” (**fudenberg1991?**), “choreographer” (Gintis 2009b) and others. [↑](#footnote-ref-53)
16. Pareto efficiency describes a state where no further improvements are possible for well-being of any individual without simultaneously decreasing the well-being of at least one other individual. [↑](#footnote-ref-54)
17. O’Connor (2019) has studied the emergence of unfairness and gendered labor division in this manner, although she has been criticized for using social categories themselves as correlation devices and not thing to be explained in the first place (Saunders 2022). [↑](#footnote-ref-55)
18. Random matching is a standard assumption in evolutionary game theory where individuals in a large, well-mixed population are paired to interact purely by chance, meaning each individual is equally likely to meet any other, regardless of their strategy. This context is important because, under random matching, the ESS depends solely on the average payoffs determined by the overall population frequencies, and strategies like cooperation typically cannot persist unless they are directly favored by the payoff structure. Deviations from random matching (assortative or structured matching) can introduce correlations between strategies, fundamentally altering which behaviors can be evolutionarily stable (Jensen and Rigos 2018; Izquierdo, Izquierdo, and Hauert 2024). [↑](#footnote-ref-57)
19. In this game, two players 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 and the cost of fighting. [↑](#footnote-ref-59)
20. The Searle’s formula involves both relations, grounding and anchoring, but for the sake of simplicity and to illuminate collective acceptance as a necessary condition for Searle’s formula I will refer to it as to anchoring. [↑](#footnote-ref-64)
21. Metaphysics differ from ontology in that the latter is a focal set of entities in a domain which are said to exist and the former is the study of broader notions of identity, causality and the like. [↑](#footnote-ref-65)
22. Although not widely known, there have been such attempts. See (**simon1962?**), (**fararo1978?**) and, most notably, Gintis (2007a). [↑](#footnote-ref-66)
23. This mirrors a longstanding debate in philosophy of mind about “folk psychology” and eliminativism, where on camp contends that “folk” concepts like beliefs and desires approximate real inner workings of the brain [#insertSource] and the other camp reduces these concepts to brain states and claims that the true inner workings can be described only in material and mechanistic terms [#insertSource]. [↑](#footnote-ref-67)
24. Hierarchical Bayes is a model with probabilistic parameters (meaning they are uncertain) which allows for hierarchical relations between random variables. [↑](#footnote-ref-68)
25. This distinction mirrors the classic one of natural and social kinds, where the former are “homeostatic property clusters”, sets of necessary and stable features (R. Boyd 1991). [↑](#footnote-ref-69)
26. Searle makes deontic powers of institutional facts dependent on language. At the same time, evidence from cognitive archaeology suggests that early hominins were able to solve coordination problems without language. Sterelny (2021) argues that social institutions (or the human social contract) conceived as shared, enforceable norms regulating cooperation and reciprocity emerged *before the advent of complex language*. Archaeologically, early hominins engaged in cooperative foraging and coordinated tool use, such as the collective extraction of high-value resources, indicating mutualistic collaboration and implicit norm enforcement in small mobile bands from around 1.8 million years ago (Sterelny 2021, 2016). Comparative primate research supports the plausibility of *pre-linguistic cooperation and rudimentary norm enforcement*, highlighting that great apes exhibit forms of mutualism and social regulation that likely scaffolded early human cooperation (**birch2022?**). Together, these findings support Sterelny’s model that proto-social contracts rooted in norms, reputation, and coordination were present in pre-linguistic societies, with language later enabling more abstract, scalable, and formalized norms as human social groups expanded. Although Searle did not conceive of institutions as of solutions to coordination problem, this is relevant from a naturalistic social ontology point of view. [↑](#footnote-ref-70)
27. Guala argued elsewhere that Lewisian conventions acquire normativity (or deontic force) in *repeated play*, so that history of play becomes a focal point for emergence of normaitvity (Guala and Mittone 2010). This view is *dynamic*, whereas the view of payoff-modifiers as repressentations of the “built-in” normativity presuppose a static view of institutions. This, along with a couple of other inconsistencies, complicates the explanation of normativity of social institutions with RiE. [↑](#footnote-ref-71)
28. As Guala notes, coordination in repeated games is achieved via modifying payoffs with parameters to increase the costs of free-riding and violation which by itself creates salience in repeated games (personal communication). However, it leaves questions of where do parameters modifying payoffs come from and how correlation of strategies and payoff modification are connected. It seems that normativity is “embedded” in RiE by default without explaining it, but it might be explained if seen dynamically. [↑](#footnote-ref-74)
29. O’Connor (2019) has studied the emergence of unfairness and gendered labor division in this manner, although she has been criticized for using social categories themselves as correlation devices and not thing to be explained in the first place (Saunders 2022). [↑](#footnote-ref-75)
30. However, it is not clear whether human players such as grazers have genuine fitness rather than utility value function. As Sterelny (2012a) suggests, there has been an evolutionary shift from fitness to utility correlated with the demographic explosion in the Pleistocene and subsequent significant decline in individual-level heritability of cultural traits, for offspring did not more resemble their parents informationally and ideologically due to the abundance of cultural information sources. He argues for model pluralism, although we are inclined to view humans agents as rather utility maximizers that fitness ones. A critical examination of the very metaphor of an individual as a maximizing agents is aptly expressed in Cédric Paternotte (2020). [↑](#footnote-ref-78)
31. As we will see in Chapter 3, normativity can be an emergent property of interaction of agents with environmental cues, which can be considered “a feature of environment”, but it would require agents to already be of higher representational grade and a system to have a separate mechanism of niche construction. [↑](#footnote-ref-79)
32. Prototype vectors refer to patterns of neural activation in the brain that represent the central tendencies of categories or concepts. Instead of relying on symbolic or rule-based representations, the brain encodes concepts as points or regions in high-dimensional neural state spaces formed through experience. Scientific explanation, on this view, involves mapping new phenomena onto these prototype patterns via neural computation and pattern recognition, emphasizing a sub-symbolic, biologically grounded account of understanding (Churchland 1992). [↑](#footnote-ref-82)
33. A real pattern, as introduced by Dennett (1991), is a structure or regularity in a system that enables predictive and explanatory power, even if it is not reducible to its underlying components. Such patterns are “real” because they provide indispensable compression of information, allowing scientists to describe systems more efficiently than by enumerating every microscopic detail. [↑](#footnote-ref-91)