MASTERARBEIT | MASTER'S THESIS

Titel | Title

Representing Institutions: How the Free Energy Principle Can Inform Social Ontology

verfasst von | submitted by

Tamás Gömöri

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 013
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Joint-Degree Middle European interdisc. master prog. in Cognitive Science
Betreut von | Supervisor:	Univ. Prof. Christophe Heintz, PhD

Abstract (German)

Diese Dissertation untersucht die Rolle mentaler Repräsentationen in der Sozialontologie, insbesondere in den Arbeiten von John Searle und Francisco Guala. Die Verwendung mentaler Repräsentationen bildet die Grundlage für die ontologischen Rahmenwerke beider Autoren. Wie ich jedoch darlege, führt ihr Verständnis von mentaler Repräsentation zu Problemen in ihren Ansätzen. Im Anschluss argumentiere ich für ein alternatives Konzept mentaler Repräsentation, das auf dem „Free Energy Principle"-Rahmen basiert, und zeige, wie dieses die Probleme in den ontologischen Theorien von Searle und Guala löst.

Abstract (English)

This thesis examines the role of mental representations in social ontology, specifically within the works of John Searle and Francisco Guala. The use of mental representations ground both authors' ontological frameworks, but, as I argue, their understanding of mental representation causes issues in their accounts. I then argue for an alternative conception of mental representation, based on the Free Energy Principle framework, and demonstrate how this addresses the issues in the ontological accounts of Searle and Guala.

Declaration of Originality

I hereby declare that this thesis is my own work unless there is clear acknowledgment of reference to others' work. This thesis has not been submitted for any other degree or professional qualifications.

Acknowledgments

I would like to thank, first and foremost, my parents who have supported me through the seemingly unending process of me getting my master's. Without their support, I wouldn't have been able to finish. I would also like to thank my friends, who listened and supported me through the difficult parts of the writing of this thesis.

I would also like to thank my supervisor, Christophe Heintz, for his patience, feedback, and commitment to see me through to the end of this 3-year process. I would also like to thank Peter Hochenauer, Valerii Shevchenko, and Ronald Sladky, who all took their time to read my work and help me to improve it.

A huge thank you to the people at the MEi:CogSci program as well, who created and maintain this program and ensure the high quality of education that it offers.

Contents

Table of Figures

I. Introduction

I.1. What Are Social Reality and Institutions?

Traffic laws are a favorite topic for debates in social ontology. Why do we have them? Why do we stop at a red light and go at green? Where and how do we cross the road safely as pedestrians? And what allows us to expect each other to stop at red and go at green? Or to not enter certain streets in certain directions? Exploring such stable, coordinated attempts to achieve certain goals together has been studied from a diverse set of academic fields, all essentially trying to understand the ontology of human social behavior. Ultimately, the main concern of social ontology is: what exactly is the nature of traffic laws and more generally, what they exemplify – social reality? (Rudder Baker 2019)

Institutions, as most commonly defined, are formal and informal rules, norms, and organizations that shape human behavior and social outcomes across a wide range of contexts, from politics and economics to culture and religion (North 1990; Powell & DiMaggio 1991; Aoki 2001). Institutions create incentives and constraints that influence individual and collective behavior, and they can either promote or hinder social progress and development (Acemoglu & Robinson 2012; Fukuyama 2011; Scott 2014).

The study of institutions is a multidisciplinary field that draws on various theoretical and methodological approaches, including historical analysis, comparative studies, network analysis, and experimental methods, among others (Brinton & Nee 1998; Dobbin et al. 2007; Greenwood et al. 2011; Powell 1991). Researchers in this field seek to understand how institutions emerge, evolve, and change over time, and how they influence social outcomes at different levels, from the individual to the global (DiMaggio & Powell 1991; Mahoney & Thelen 2010; Ostrom 1990; Zucker 1977). This endeavor has produced a wide range of perspectives as to the nature of institutions, their building blocks and most basic mechanisms, and has resulted in distinct intellectual camps who each approach institutions according to their effects. Anthropologists from the lens of cultural change and stability, economists to examine decision making processes through game theory, political scientists to examine the role of institutions in government, policy decisions, international relations – the list goes on. These perspectives, do however share some similarities as well.

The anthropologist João de Pina-Cabral (2011) explains that institution, as a concept, has a long history and has come to be associated with certain societal forms – governments, marriage, corporations, etc. This is due to a conception of institutions as "hierarchical systems bound by explicit, rational, written codes" that dominated Western philosophy and the social sciences throughout the 20th century (p. 478). This has implications for the way we think about society: "as a formally coherent system that imposes itself upon individuals" (p. 479).

He goes on to say that our current conceptions of institution constitute "semantic slippage". We combine two definitions – one from Marcel Mauss and one from Max Weber. De Pina-Cabral refers to Mauss' theory as "minimalist": it characterizes institutions simply as "public rules of actions and thought" (p. 482). Weber's is "complex": "compulsory organizations" (often translated as institutions) have a "rationally operating staff" out of which some person(s) carry into effect the order governing the organization. Furthermore, they impose this order in instances of action conforming to certain criteria. The two forms of "semantic slippage" arise from the combination of these definitions. First, the combination of these two definitions implies an evolutionist perspective. Institutions start simply as "public rules of action and thought" that can exist in non-literate cultures and eventually solidify into complex, modern-day "compulsory organizations" that can only exist in civilized, developed societies (Weber's primary example of a "compulsory organization" is the State). Second, these two definitions seem to refer to two different ideas: Mauss' very general definition is about institutionalization, while Weber's actually describes instituted institutions, or specific instances of institutionalization.

While de Pina-Cabral's summary of thinking on institutions is not comprehensive or exhaustive, it is a crucial backdrop in the current, lively debates about the ontology of institutions. As he highlights, most theorists' examples and formulations of institutions have an underlying theme – they take the social structures of Western, industrialized societies to be the norm – what he would describe as "a contractual model of society" operating "with notions of self that are individualist, logocentric and mentalistic" (p. 491).

It is important to keep in mind that the definition de Pina-Cabral outlines has deeply influenced the thinking of social ontologists, but their own definitions vary greatly by what they think are the underlying mechanisms behind institutions. There are a multitude of perspectives on the nature of institutions and attempts to formulate adequate theories have been made numerous times in a variety of academic fields, including anthropology, sociology, economics, and political science. DiMaggio (1998) argues that the newest theoretical frameworks within these disciplines can be called neo-institutionalist, as they do not embrace individualism as a starting point for their explanations, and instead stress collective phenomena and their causal roles, such as rules and strategic interaction. This is a contentious point, as not all neoinstitutionalist frameworks which highlight the importance of collective phenomena also reject their reducibility. Searle (2010), for instance, while maintaining that collective intentionality cannot be reduced to individual intentionality, does not claim that there exists a detached social reality outside of human minds.

Out of DiMaggio's three main categories of neo-institutionalisms – Rational-Action Neoinstitutionalism (RAN), Social-Constructionist Neoinstitutionalism (SCN), and Mediated-Conflict Neoinstitutionalism (MCN) – I will be examining work in both RAN and SCN, represented by work from Francesco Guala and John Searle respectively. I will examine whether their views on representations truly provide an *ontologically* plausible framework and examine challenges and alternatives based on an interactionist view of institutions and the Free Energy Principle.

I.2. The Issues With Mental Representation in Searle and Guala's Accounts of Social Ontology

Current attempts by Searle and Guala to formulate theories within social ontology embrace a naturalist framework. Both Searle and Guala build their accounts on contentful, feed-forward, and brain-based mental representations (hereinafter, traditional), which ground their theories and tether them to a naturalist framework. Contentful means that any representation carries with it some content about the outer or inner world. Feed-forward means that such representations are created from the bottom up: they are axiomatic, meaning they stem from perception and are the essential building blocks of increasingly complex and abstract thinking. But overall, the direction of information processing is from environment to organism, lower levels to higher. And finally, brain-based refers to the idea that all representations are instantiated by the brain and can be found exclusively within people's head. The body and environment are involved in creating representation, but do not directly constitute it.

Both authors claim that in the end, all social phenomena can be understood as the social distribution of mental representations within a community, which exist within individual brains, and the behaviors based on these representations. They also claim that such mental representations can be further deconstructed to physical phenomena. It is this theoretical commitment that ultimately leads Searle and Guala to rely on mental representations to be the linchpin between social and individual, physical and mental, nature and nurture. I will detail what sorts of assumptions are implied in the work of both authors about mental representation and how their conceptions have been challenged in recent years by work in philosophy of mind and neuroscience.

For Guala, mental representations are the basic building blocks on which he can build an account of institutions. For Searle, mental representations are the linchpin that connects his deontology to a naturalist conception of mind and allows him to construct a rule-based account of institutions. In both, the commitment to traditional mental representation creates inconsistencies in their accounts. This theoretical commitment creates 4 problems for potential social ontological conceptions, which will be explored in detail in the thesis:

1. They burden mental representations with the task of connecting the individual to the social. Somehow, individual actions and institutions influence each other. Can a traditional representation truly account for this connection? No: for Searle, there are no incentives for following rules; their power to influence behavior comes from collective intentions, a type of shared mental representation. His account of representations thus don't actually explain the mechanism by which they exert power over individuals because they are built on intentionality, not incentives. For Guala, there is no explanation for how cultural change happens (nor does he ever attempt to explain why it exists) – his explanation hinges on evolution endowing humans with needs and desires in the form of representation that create the basis of the decision-making within institutional settings. Thus traditional representations create stability, but do not explain how novelty and unique solutions are created to coordination problems with many possible solutions.

Looking at individuals through the lens of evolution can present them as overly stable, but we once again lack a connection to the dynamically changing nature of social life.

2. They burden mental representations with the task of connecting the mental and the physical. Traditional representations begin and end with the mind/brain. Both authors conceive of a mechanistic world which is understood through physics and chemistry, thus the mind and brain are also mechanistic, but work with representations. So how does a mechanistic and physical world and brain give rise to mental representations? In other words, how do representations acquire content? For Guala, the process rests on the idea that evolution has shaped our minds to notice similar patterns in the environment, thus allowing us to create shared mental representations that allow for coordination. The entire process is quite automatic and thus does not give any ways in which new solutions and adaptations are generated – if we all notice the same patterns in the environment, do we only change our minds when the environment changes? For Searle, the process of generating new social functions – such as declaring paper to be "money" – creates social reality, but he gives very little explanation for how the physical world influences social function creation and maintenance. Can anything be declared money, for instance, or are there physical limitations to what can reasonably be used as money?
3. They pigeonhole social reality into either top-down or bottom-up processes. For Guala, and determine actions, which are then aggregated to create institutions. Our minds detect patterns in the environment and create mental representations that reflect these patterns. It seems to be implied that our institutions are bottom-up, since our environments (whether physical or social) determine what we represent and how we act. For Searle, insitutions are fundamentally normative and make participants act in ways they otherwise wouldn't want to. This is top-down approach ignores the environmental incentives that constrain and drive the sort of behaviors we engage in. Why would participants subscribe to institutional rules without individual incentives? (To take money as an example again, it would be highly impractical to declare elephants as currency, so we don't – Searle spends very little time focusing on these kinds of practical constraints).

4. They use mental representation as the main explanation for why human cognition is separate from animal cognition. Searle appeals to special, collective intentions – a special type of representation – to do the job of allowing humans to make social reality in ways unlike animals can. We can represent collective possibilities for action, allowing us to "go beyond" the basic facts of our environment and instead create novel social functions. Guala, on the other hand, conceptualizes social coordination through game theory, which can be used to describe both animal/plant and human coordination. Therefore, Guala must somehow distinguish animal/plant coordination from human social coordination, otherwise animals would also have institutions. He thus appeals to mental representations as the mechanisms which separate animal behavior from human. Animals do not represent, but are simply simtulus-reaction machines, while humans do. Their behavior can be better described through stimulus-representation-reaction.

I.3. The Free Energy Principle as an Alternative to Traditional Representation

So what could be an alternative to the traditional account? In contrast to the traditional representations that are found in Searle and Guala's accounts of social ontology, the Free Energy Principle (FEP) framework can potentially be leveraged for a new cognitive basis for social ontology. It reformulates what representations are, by extending them beyond brains, conceiving of them instead as inferential models that attempt to predict the environments around them. It conceives of cognition through such inferential models, which are regulated by several important interactions. Free energy is an estimation of the constant misalignment between the distal causes of the sensory states (the environment) and the inner states of the biological system, and its minimization can work both in the direction of sensory states (perception) or toward active states (action). Over a much longer timescale, it is also responsible for the shape of the organism and its niche (environment). Through attunement, the inner states of the biological system are maintained within a viable set of physiological bounds – also known as allostasis (Sterling & Eyer 1988; Sterling 2012). This is fundamentally different from the sort of minds Guala and Searle make appeals to, as the FEP sense-making does not attempt to represent reality, but rather seeks to engage with it through action. As such, representations are not reflections of "what's out there" or "what do I believe", but simply guesses as to what potentially may keep the interaction between organism and environment most stable – whether new actions or updated inner models are required for this stability (Friston 2010).

So a social ontology based on the FEP solves the problems posed by one based on traditional representations: the individual and social, animal and human, and mental and physical are all joined under the principle of minimizing free energy. The FEP framework can be considered interactionist, as it describes cognition as fundamentally serving to keep organismic and environmental dynamics stable – both are vital in the process. It does not presume that representations are entirely within the mind, but exist throughout the body as well. Thus, the FEP lays a groundwork for representations that are not entirely mental; this differs starkly from the representations present in Guala and Searle. The FEP also outlines a holonarchical understanding of organismic sense-making: there are top-down and bottom-up processes, all of which serve allostasis (extended to the social, both institutional- and individual-level processes matter). And finally, free energy minimization is not a unique endeavor of the human mind, but is also done by living systems as simple as single cells.

There is some preliminary work on using the FEP as a theoretical foundation for social reality (De Bruin, Van Elk, & Newen 2012; Constant et al. 2019; Veissière et al. 2019; Ramstead, Constant, & Badcock 2019). In such an account, the integration and differentiation of different parts of the systems (from cells to institutions), their functions, and their interaction is accomplished by the minimization of free energy. This reflects the FEP's basis in dynamical systems theory, which posits that structure and function mutually cause each other and thus the interaction between the two is the focus of attention (Bateson 1972; László 1972; Macy 1991; Stroh 2015). This is, however, quite difficult to make sense of because the FEP struggles to define clear limits to the process of free energy minimization – it runs the risk of either confining cognition by equating it to organismic models (the environment is not strictly involved in cognition) or it expands cognition to include the environment, in which case representation and cognition become unwieldy and difficult to delineate (does the environment outside of organisms think, or represent?). This problem is not easily resolvable.

I.4. Main Argument and Thesis Structure

In this thesis, I will outline some existing issues by looking at two current conceptions of institutions – those of Searle and Guala – and, based on principles found in the FEP framework, outline important considerations for a cognitively informed social ontology. This cognition is based on an interactionist, embodied, and enacted view, in which representations – whatever they may be – go beyond minds and potentially individuals; they are systemic. I will suggest that social ontology could benefit from considering how the FEP framework enriches its theoretical basis to better accommodate the four issues outlined in section I.2. The object of social ontology can be thought of as defining and delineating the social – to answer "what genuinely and nonredundantly exists," and the FEP provides a framework for how livings systems come to exist and stabilize (Rudder Baker 2019, p. 1). It is not a big stretch to use the logic behind the FEP in living systems and extend it to social systems. Throughout the thesis, I will explore what problems exist in Searle and Guala's representations and the issues these pose for the frameworks. I will then give details about the Free Energy Principle and discuss the work that has been done to connect it to social cognition and social ontology. I will also highlight some examples of how a FEP conception of social reality can resolve the issues highlighted in traditional accounts. I will then highlight some inconsistencies in the FEP framework that may pose challenges for its use as a basis for representations and for understanding social ontology.

In conclusion, I will outline that, much like how the FEP outlines an understanding of cognition which is thoroughly dynamic and action-oriented, it makes sense to think of social ontology as action-oriented and systemic. Action-oriented puts emphasis on the idea that humans as biological systems exist to act on their environments (and themselves), not to represent them. Their successful engagement with the environment is primary, their understanding of it is always co-determined by the success of their actions and built to drive successful actions. Such a systemic view can serve as a way to find new potential avenues for research by rethinking how we approach institutional phenomena. It provides challenges to current conceptions of mind, cognition, and the social, and provides an alternative explanation for representations, while allowing for a more nuanced and dynamic view of institutions. Interactionism and the FEP posit that mechanism and function are always co-determining: the ontology of the social is the process of interaction itself, not the behavior of its constitutive parts. So what would it look like to participate in the interaction, rather define it?

II. A Critique of Traditional Accounts of Social Reality

II.1. Outline of the Searlian Account of Social Reality

John Searle builds his account of social reality on the ability of humans to represent the world around them in a unique way – with the ability to create a special kind of collective intention, which allows them to turn “plain” objects, acts, agents, and give them a function that is fundamentally social (Searle, 2010). Collective intentionality[1] has gained a lot of traction as a theory for explaining the connection between social institutions and basic human social cognitive abilities. There are several different arguments maintaining that the existence of collective intentionality is necessary for the existence of human sociality and institutions. Searle does not think of collective intentionality as conflicting with naturalism or mental representation. In fact, he develops his account of institutions by keeping in mind that intentions, which are collective, do not leave the skull of individual human beings – they are still mentally represented. As we will see later, this commitment causes issues in justifying how collective intentions actually influence individual behavior as there is not much emphasis on an explanation for what incentives drive humans to adopt collective intentions.

A human institution, according to Searle, is then a set of constitutive rules. But what exactly are constitutive rules? To understand them, we must first understand that they exist to create status functions. Status functions are, according to Searle, the most fundamental and most unique feature of human sociality:

> The distinctive feature of human social reality, the way in which it differs from other forms of animal reality known to me, is that humans have the capacity to impose functions on objects and people where the objects and the people cannot

[1] In philosophy, intentionality refers to the capacity of minds and mental states to be directed toward, represent, or be about objects, properties, or states of affairs. When mental states have intentionality, they are representations or possess content. Additionally, when a person uses language, symbols, or images to communicate their thoughts, these tools also carry intentionality, as they represent the content of the speaker’s mental states (Jacob 2023).

> perform the functions solely in virtue of their physical structure. The performance of the function requires that there be a collectively recognized status that the person or object has, and it is only in virtue of that status that the person or object can perform the function in question. (p. 7)

Status functions are like the building blocks of Searle's institutional reality and are used in constitutive rules that take the following format: X counts as Y in context C. For the institution of the United States dollar, for example, a piece of paper with certain physical properties (X) counts as currency (Y) in the United States (C). Through these constitutive rules, humans are able to construct institutional reality, which can be described by institutional facts – facts that only exist within human institutions. They are unlike basic or brute facts, which Searle defines as being explainable through physics, chemistry, and other natural sciences.

Searle's institutional reality is built fundamentally on intentionality. It is important to highlight that Searle conceives of minds as fundamentally intentional and conscious, and that this combination is what leads to brains creating representational content (Searle, 1983). Unlike computationalists, Searle claims that brains represent the world directly by virtue of their biology. Searle does not view the mind as computational symbol manipulation; he does, however, conceive of the mind as having representational content and at the same time being entirely based on physical processes. Contentful representations pose issues, even if they're not computational, but this will be discussed in later sections. For now, the first issue with this traditional representations is that we are left wondering: how exactly is it that mechanistic, physical processes bring about representational states?

II.2. Mental Representations as Connection Between Mental and Physical for Searle

This is where we see the first issue of the Searlian account. The implication above is that brute facts at some point somehow transition into institutional facts qua collective intentionality. Since intentionality itself implies being about or representing a state of affairs, this means that representationality is at once the ground on which human mental (and social) lives are built and its unique separateness from brute facts. Searle insists that his social ontological framework rests firmly within naturalist assumptions – namely that all collective attitudes exist in individual minds and are reducible to neural activity, biochemistry, physics – but it is unclear how the function of intentionality as a separation between brute facts and mental facts can work as a naturalist ontological framework. If representation can be reduced to simple brain chemistry, at what point does it acquire its unique, mental nature? This is a circular argument: representations connect the individual's mental world and social realities by allowing for the creation individual and collective intentions – and then status functions – while at once being entirely reducible to brain chemistry. Lynne Rudder Baker (2019) puts it this way:

> Since ontology concerns existence, not knowledge (Searle 2010, p. 18), the epistemic objectivity of the social world has no bearing on ontology. On Searle's view, what makes something social is an epistemic matter of our conferring a status on physical phenomena. A cocktail party is a gathering of people that we count as a cocktail party. So, when Searle says that the social world is "ontologically subjective," the word "ontologically" is just a façon de parler; in light of Searle's general ontology, what is ontologically subjective has only epistemic import, no ontological significance at all. Otherwise, ontology would not conform to physicalism. Hence, given Searle's conditions of adequacy for a theory of social reality – a theory of social phenomena in a world that is entirely physical – there is logically no place for social phenomena in ontology. Putting social phenomena in ontology would either render the ontology redundant, or would abrogate Searle's physicalism. (p. 4)

Searle implies that intentionality (mental representation) creates a new kind of fact, which is more than basic or physical, but also somehow fits into a naturalist framework where everything is reducible to chemistry and physics.

II.3. Mental Representation as Connection Between Individual and Social

Searlian institutional reality is, despite this circularity, built on status functions, which require prior intention, intention-in-action, and collective recognition to exist. All of these different mental states can be boiled down to traditional mental representations and also do the explanatory work of explaining how individual mental states and collective mental states motivate individuals socially, while also remaining entirely within individual minds. This is the second issue in the Searlian account, where mental representation must somehow contain all sorts of mental states within individual minds, but also account for how social normativity can then influence individual minds. So how do these mental states bring about a functional social reality?

Prior intention is what causes an agent to perform an action and exist before the action itself is completed, while intentions-in-action are the intentional components of action as they are performed. This is the mechanism by which Searle tries to explain the integration of I-intentionality and we-intentionality: collective prior intentions can arise from institutional facts, yet the content of their corresponding intention(s)-in-action can be entirely I-intentionial. For instance, voting in a board meeting requires a we-intentionality of voting, which implies collective cooperation since it holds no meaning outside of collectively recognized institutional facts, and is carried out by an intention-in-action of hand-raising, which in and of itself is devoid of institutional facts or we-intentionality. This way, individual actions become institutional fundamentally through the quality of their representative content (i.e. am I voting or am I simply raising my hand).

The third requirement for status functions to exist, collective recognition, depends on a weaker form of collective attitude. Here again, we see the split between social representation (we-intention) and individual I-intention. Collective recognition is the attitude that allows for all parties involved in cooperation to accept and identify institutional facts. For instance, before being able to take part in a store purchase, the seller and buyer must collectively recognize pieces of paper as money, which can be used in the exchange of goods and services. As mentioned above, Searle argues for the irreducibility of we-intentions (intentional states that are aimed at cooperation) to I-intentions. This means that, in order for cooperation to happen at all, all parties involved must have the ability to have shared intentional states, since individual I-intentions do not add up to collective we-intentions. Not only is going for a walk together different from walking with someone at the same time by chance, but the idea of walking together cannot exist without collective intentionality. In other words, social reality cannot be constructed without representation.

The idea that in order for cooperation to exist, all parties must have shared intentional states points to the crux of the problem. Mental representations can be reduced to brute facts – brain chemistry – and intentional states can be reduced to mental representations. In other words, all agents must have the same exact mental/brain state in order to cooperate. While in theory this could be possible, but seems highly unlikely – if we were able to share such deeply analogous intentions, disagreements, mistakes, miscommunications, among other mismatches in cooperation would likely not exist. There is also increasing evidence from studies in neuroscience pointing to a brain that is continually constructing its internal states actively, and is influenced by both bottom-up constrains and top-down/outside influences like, cultural, emotional, and physical factors (Karmiloff-Smith 1992; Feldman Barret 2017; Astle, Johnson, & Akarca 2023; Karmiloff-Smith, Thomas, & Johnson 2023).

II.4. Interlude on Collective Intentionality

Searle advocates for the existence of collective intentionality[2] and its essential role in the existence of human institutions, and is a point also made by Raimo Tuomela (Tuomela 1995, 2007; Hindriks 2003). Another version of this argument exists, which states that individual agency does not account for certain forms of cooperation – namely, large-scale charity efforts with thousands of anonymous individuals – which thus require some form of shared intentionality (Hindriks 2003; Sugden 1993). Another argument commonly made for collective intentionality is the difference between the content of intentional states for a group and for an individual. Margaret Gilbert's (1989, 2009) view posits that collective intentions form a plural subject that transcends individual mental states, representing a kind of group agency or collective commitment. Michael Bratman (1993, 2014), on the other hand, sees collective intentions as interdependent but still fundamentally individual mental states. For him, the coordination and cooperation between individuals do not require a collective entity beyond their individual intentions.

Collective intentionality is a concept that describes the ability of two or more individuals to share a common goal or intention. According to Bratman (1993), collective intentionality requires the ability to communicate with others and to coordinate actions towards a shared goal. Gilbert (1987) argues that collective intentionality is a fundamental aspect of social life, as it

---

[2] Collective intentionality is the idea that humans can have collective intentions, or "we-intentions," which are not reducible to individual "I-intentions" (Searle 2010). Shared intentionality is a related concept which encompasses the uniquely human ability to share psychological states (intentions, goals, attention) with others in cooperative contexts (Tomasello et al. 2005)

enables individuals to engage in joint actions that they would not be able to achieve on their own. Furthermore, Searle (1990) posits that collective intentionality is even essential for the creation and maintenance of social institutions.

The study of collective intentionality has important implications for understanding social phenomena such as cooperation, and social norms/institutions. According to Tomasello et al. (2005), the ability of individuals to share a common goal or intention is necessary for the emergence of cooperative behavior. Moreover, shared intentionality is also necessary for the emergence of social norms, which are rules that govern social behavior. Social institutions, such as governments, legal systems, and markets, are also based on collective intentionality, as they require individuals to share a common understanding of their roles and responsibilities. Despite its importance, the concept of collective intentionality is still the subject of much debate and research. Some scholars have raised questions about the nature of collective intentionality, and whether it is truly distinct from individual intentionality (Gilbert 2014; Guala 2016).

II.5. Mental Representation as Division Between Humans and Other Living Things

Collective intentionality can be thought of as a suite of cognitive mechanisms unique to humans, falling under the umbrella term “collective” or “we-intentionality”. They are the cooperative mechanism that allows humans to “unlock” the social realm, to create rules and work together in a way which no other living thing can. Tomasello & Carpenter (2007) define shared intentionality as “collaborative interactions in which participants share psychological states with each other”. (p. 121) To give an example, for the activity of walking with someone, shared intentionality is the ability of both parties to walk *together* – this is very different from two people who individually have intentions to walk, but do so at the same time and place, next to each other. According to the proponents of collective intentionality, it cannot be created by adding up individual intentional states – the individual intentions of two people to walk in the same place, at the same time, next to each other do not automatically add up to the activity of walking together (Searle 2010). Instead, these intentions result simply in walking at the same time, in the same place, in the same direction without real cooperation, by accident.

When viewed this way, collective intentionality can be construed as a special type of mental representation: a we-intention. These mental representations, where individuals share a cooperative intention, are distinct from I-intentions, which only apply to a single individual. Elisabeth Pacherie (2005) summarizes their uniqueness:

> The idea then is that to account for cooperation we have to introduce a specific type of mental states: we-intentions. But in what sense are intentions special and in what sense can they be said to imply cooperation? Prima facie, three possibilities are conceivable. First, it might be that their contents are what makes we-intentions special and thus that the dimension of cooperation is linked to specific features of these contents. Second, what makes we-intentions special could have to do with the type of entities they can be attributed to. The third possibility is that rather than the contents or the possible bearers of we-intentions, it is the psychological mode itself — i.e., the fact that the psychological mode is that of we-intending instead of I-intending — that implies the notion of cooperation. (p. 156-157)

For the first two possibilities, Pacherie explains that they do not work for Searle, as he conceives of a social reality that is present entirely in individual minds – no content can make we-intentions separate from I-intentions, and they can only be attributed to individuals, not entities with a group mind.

The third possibility, the psychological mode, is what sets we-intentions apart for Searle. To put it into a concrete example: pushing a car alone while mistakenly thinking that somebody else is helping, is a case of mistaking our intention to be attributable to a "we", therefore, mistaking the type of intention that one has (an I-intention as opposed to a we-intention). As Pacherie (2005) points out: "it seems that, contrary to what Searle claims, his analysis is not consistent with internalism. An individual cannot properly be said to have a collective intention unless other individuals actually share his intention" (p. 161). The proponents of we intentionality claim that it is not only the building block of social and institutional reality, but that it is a necessary precursor to all levels of human sociality (Vromen 2003). However, from Pacherie's analysis, it is already apparent that certain theoretical assumptions about the nature of collective intentionality as a mental state pose problems for its viability as a basis for social ontology.

Searle is closer to Bratman's point of view and stresses that an individual as a team member must derive her goal from the group's intentionality, but maintains that the we-intention of the group nonetheless only exists in individual minds. For example, a construction worker would derive her individual intention of hammering a nail from the group intention of building a house – where the group would consist of other construction workers, a contractor, an architect, the future homeowners, etc. Bratman, however, does not align with Searle that special we-intentions are used in cooperation; instead, cooperation happens through the planning of individual intentions that can mesh and come with mutual responsiveness from all involved.

Tuomela and Margaret Gilbert make a different form of this argument: a group may have a certain intention (like a belief) that does not correspond to the intention of an individual as a private person (Hindriks 2003). For instance, our construction worker might disagree with how the house is being built privately, but that does not stop her from acting in accordance with the group intention of building the house in such a way. There must be a collective intention, then, which she adheres to, that can be distinguished from her private intention. This does not necessarily mean that her decision to adhere to the group intention is *caused* by the existence of the group intention, but that there is a difference between her private and group intention and that without the group intention, her individual action does not add up to anything that can be considered social.

II.6. Mental Representation as the Basis for Top-down Institutions

A third reason that collective intentionality is so central to the Searlian account is that they are the basis for what Searle calls "deontic powers" (Searle 2010). Status functions carry deontic powers, which are essentially reasons for action beyond self-interest. In other words, collective intentionality, a mental representation, gives institutions the ability to exercise power over individual actions and decisions. Such deontic powers are thus a top-down form of causality. The example Searle gives us is the ability to conceive of boundaries, like those of a village. Boundaries that are not physical, but socially conceived, require the collective recognition of a group. This collective recognition then gives normative incentives for individual group members to act certain ways (like stay within or outside of a boundary). So what is the issue with this normativity?

Searle highlights that collective intentions are not only precursors to any idiosyncratic, cooperative activity, but also serve as the basis for normative behaviors. In other words, not only are cooperative interactions built on collective intentions (two people going for a walk must have a shared intention of "we are going for a walk"), but normative, rule-based interactions are also dependent on them. For him, the only way to understand institutional facts, which go beyond physical facts, is to characterize them as normative – collectively recognized by a certain group. Deontic powers, arising from the collectively recognized status functions given to different objects, events, people, etc., give members of the group reasons for action beyond their self-interest. According to Searle, this is ultimately what gives institutions their institutional nature.

This is where we see the fourth issue with a social ontology based on traditional representation: if intentions (which are mental representations) can be recast as collective intentions (also mental representations), then deontic power arises to make institutional participants behave in accordance with the rules. This leap is an extremely big one, as Searle claims "conventions are arbitrary, but once they are settled they give the participants a right to specific expectations. They are normative" (Searle 2010, p. 87). It is overly "top-down" in the sense that arbitrary norms based in traditional representation have the power to influence individuals, despite the very nature of the mechanism of representation – non-aribtrary and reducible to physics and chemistry. It is odd to assume that status functions can be given arbitrarily and that normativity arises from collective acceptance, as if environmental circumstances and cognitive limits have no import on how institutions are shaped. This point is heavily contested by rational-action neoinstitutionalists, who posit that the material and mental conditions that allow certain institutions to exist are ontologically much more central than the ability to create rules and status functions.

II.7. Introducing Equilibria

An alternative approach to institutions that does not involve collective intentionality views institutions as solutions to coordination problems. Searle's account of institutions can be understood fundamentally as a rule-based account of institutions (Guala 2016). Searle advocates for constitutive *rules* being the defining mode in which humans endow their physical environment with institutional structure. If we collectively recognize certain rules, and these rules have deontic power, they will then motivate us to act in accordance with the rules (or sometimes, to intentionally break them – but whether we knowingly break or abide by the rules, we still recognize them as existing). In opposition to this, the equilibria approach to institutions treats institutions as stable forms of behavior that arise from coordination problems (Guala 2016). These behaviors remain stable because individuals have no incentive to deviate from them. Once again, as will be discussed later, Guala commits to a theory of institutions where mental representations are stable and traditional, as they serve as the source of motivations, beliefs, and attitudes that drive behavior within coordination problems. In essence, mental representations are the source of the incentives for behavior around which institutional equilibria emerge. This theoretical commitment leads to an unrealistically stable view of institutions (Akiyama & Kaneko 2000) and cannot account for how specific aspects of institutions emerge or how such aspects then influence the beliefs, attitudes, and desires of the agents that make them. As we will see, the equilibria account is heavily reliant on mental representations: they must enable the creation of correlational devices which innovate and stabilize behavior, distinguish human institutional equilibria and animal equilibria, act as the bridge between evolutionary history and idiosyncratic institutional cooperation, and provide the basis for why individual incentives drive institutional behavior.

In his book, *Understanding Institutions* (2016), Francesco Guala advocates for the unification of these two seemingly disparate approaches. Much like Searle – who identifies status *functions* in the form of constitutive rules (X counts as Y in context C) as the mechanism by which everyday physical objects, people, places, etc. become institutional – the rules-in-equilibria approach takes as a given that institutions perform some function for a given social group. The function is the facilitation of coordination through specific solutions that abstract away from the myriad ways in which a goal can be achieved, which aligns with the interests of the group's individuals (Guala 2016). Guala, however, insists that it is not necessary to think of institutions as being dependent on collective intentionality. Instead, certain aspects of institutions-as-rules and institutions-as-equilibria must be combined to adequately describe the nature of institutions.

The combination of these two approaches involves a certain amount of game theory and some ideas about convention. The rules part of the approach gives us a certain understanding of what actions are appropriate in different circumstances:

> The institution of private property, for example, regulates the use of resources by indicating who has access to them. The institution of money regulates the use of paper certificates in economic transactions. And the institution of marriage regulates the behavior of two or more individuals who pool their resources to raise kids, manage property, and help each other in many different ways. (p. xxiv)

Just having rules, however, does not guarantee that people will follow them. The equilibria part is necessary because it provides a "system of incentives and expectations that motivate people to follow the rules" (p. xxiv). Different problems of coordination have different incentives and expectations for players, but problems that can be resolved through equilibria are special cases: "no player can do better by changing her strategy unilaterally. If the others do their part in the equilibrium, no player has an incentive to deviate" (p. xxv).

II.8. How Mental Representations Ground Incentives, Bridging Mental and Physical

Institutions can thus be thought of as answers to different types of coordination problems. Rules, whether informal or formal (e.g. the rules of friendship vs. the rules of how a court proceeding must go), must be backed up by incentives for the individuals following them so that they can be stable. The way in which rules and equilibria are combined in Guala's theory, is through correlated equilibria, which is rooted firmly in game theory, David Lewis' ideas on convention, and Schelling focal points. All in all, Guala makes the argument that whenever there are conventions, they arise from shared mental representations of focal points in the environment, which are used as correlation devices to create correlated equilibria. Once created, the correlated equilibria are simply maintained because no unilateral deviation is profitable for participants. Institutions (rules, norms, signals) make compliance a best action given others' expected compliance (and sometimes given enforcement). Thus, we have a mechanistic description of how mental representations of the environment (Schelling focal points) lead to incentives. In other words, mental representations once again act as a bridge between the physical and the mental.

For a given game, a Nash equilibrium is reached by augmenting the original game with the addition of novel strategies. A Nash equilibrium happens when the players of a game do not change their strategies even after considering what the other player is most likely to do. The novel strategies are conditional, based on an external event that is not part of the original game. They are conditional, because the player's action depends on the occurrence of the event (if "X" then do "Y"); these events are correlational devices.

A simple example of this would be an intersection where two cars meet, which can be described as a hawk-dove game. Let's assume that two cars approach an intersection and both slow down to figure out who can pass first. Car A passing before car B constitutes a bigger payoff for car A (since it gets to where it is going faster) and a smaller payoff for car B, but overall, they avoid collision; this scenario can be represented as (2,1). Following this logic, the alternative would be (1,2), where car B passes first, (1,1) would be if both cars stop and wait indefinitely thinking the other will pass first, and (0,0) would be when both accelerate at the same time, resulting in a crash. Knowing that the other car is likely to be rational and wants to avoid crashing, it makes sense for a car to choose to go first. Therefore, the two Nash equilibria are (1,2) and (2,1).

This is where a correlation device comes in: in this case, a streetlamp. The streetlamp's colors are external to the original game, but they add new strategies and payoffs to it. This way, if both cars follow the rule imposed by the lamp (go if green, stay if red), they both have a 50% chance of getting green to go and 50% chance of getting red and having to stay. They thus end up in an equilibrium in which the expected utility is equal to the equilibria of the original game, (1,2) and (2,1), but their payoffs are equal and they are guaranteed not to crash.

	Go	Stay	Go if green, stay if red	
Go	$(0,0)$	$(2,1)$	$(1,\frac{1}{2})$	
Stay	$(1,2)$	$(1,1)$	$(1,\frac{3}{2})$	
Go if green,	$(\frac{1}{2},1)$	$(\frac{3}{2},1)$	$(\frac{3}{2},\frac{3}{2})$	stay if red

Figure 1. The different payoffs of the original game for two cars crossing an intersection can be seen in the box with the bold outline. A correlation device (the streetlamp) provides new strategies and payoffs. A correlated equilibrium is reached when both cars choose to follow the streetlamp and have the payoff ($\frac{3}{2},\frac{3}{2}$).

The overall argument Guala puts forward is that when some problem requires cooperation, these sorts of correlated equilibria come about naturally as a result of humans' mind-reading – we are excellent at inferring others' mental states so that we may have a clear picture of what our options are for action, given what others think – and the environment. The equilibria do not require constitutive rules. Searle's constitutive rules, he argues, can be reduced to regulative rules – rules that simply regulate behavior instead of creating it, such as our streetlamp example. There is no need for collective recognition of rules that create new behavior, they arise naturally from a combination of human mind-reading and the environment. To explain how these rules come into existence, he combines David Lewis' ideas on convention with Schelling's focal points. Where there are problems of coordination, repeated behaviors that are motivated by payoffs in the coordination become conventions. We are able to infer others' mental states from their actions and adjust our own behavior to fit conventions. Correlational devices, on the other hand, arise from focal points in the environment that everyone converges on due to shared human cognition, which is why different people can eventually develop commonly accepted rules that regulate their behavior.

II.9. Mental Representations in Guala: Bottom-up and Connecting Social to Individual

Guala also relies on the process of: traditional mental representation, which reflects the relevant parts of the environment, which leads to incentives, which lead to convention. He does not flesh out the details of how exactly he conceives of mental states and representations in the way that Searle does, but how he uses them points to a very similar assumption. The mind encodes what the senses perceive (Schelling focal points), which provide the building blocks for incentives, correlational devices, and cooperation. This process, unlike Searle however, is entirely bottom-up: evolution and the environment shape what individuals find relevant and thus, the implication seems to be that we all zero in on the same relevant focal points in the environment, allowing us to cooperate. In this way, the representations are bottom-up (meaning the source of cooperation is from shared evolutionary history, mind-reading, and ques from the environment, not deontic powers – there is no "downward causality" from institutions, they arise from the decisions and actions of their participants) and they once again connect the individual to the social. By traditional mental representations in individual minds we are able to coordinate socially.

Institutions that are built entirely bottom-up fail to address how cultural values shape individual values and, as highlighted by Przeworski (2004), they fail to explain why institutions exist and matter. But Guala claims that constitutive rules – the backbone of deontic power – are not necessary. To demonstrate, he provides an example of how the institution of property develops between two tribes, the Dinka and the Nuer, who move into the Sobat Valley to graze. There is a river that provides a natural separation for the two tribes, so they each settle on the North and South side of the river. Over time, the river dries up, leaving only a sandy riverbed stretching between the territories grazed by Nuer and those grazed by Dinka. This opens up the possibility for the two tribes to cross over into each other's grazing areas, which could lead to violence or war. What has arisen is a coordination problem that can be described by the same hawk-dove game as for the cars crossing an intersection. When seeing an empty piece of land, if both tribes graze, violence breaks out: (0,0). If one tribe grazes and the other defers (2,1) or (1,2), and if both defer fearing violence, (1,1). Since the Nuer and Dinka have used the river up until now as a border, it is a salient focal point in the environment (at first, this was because of its physical characteristics – it was difficult to cross). It is possible that a regulative rule thus develops for the tribes that can be followed in order to expand the strategies available in their coordination game: if the new piece of land is North of the riverbed, then Nuer graze, if it is South, then Dinka graze.

Guala then explains why this is significant: we could now call land North of the riverbed Nuer property and South of it Dinka property. This comes in the form of a regulative rule: "If a piece of land is Nuer's property*, then the Nuer graze it" (p. 64). According to Searle, property is an institutional term would require a constitutive rule, something like this: land (X) counts as Nuer property (Y) North of the riverbed (C) (Searle 2015). Guala rearranges it to show that it can be turned into a regulative rule: If North of the riverbed (C), the land (X) is Nuer property (Y), and if it is Nuer property (Y), then the Nuer graze it (Z). Since the regulative rule at the beginning stated land is Nuer property if Nuer graze it, the Y term in the previous formulation is redundant. It could be expressed as: if North of the riverbed (C) then Nuer graze it (Z). This means that institutional terms do not describe a new level of reality – institutional facts. They are, in fact, eliminable:

> That institutional terms are elimin*able*, of course, does not mean that they must be eliminat*ed*. On the contrary, the introduction of theoretical terms has an important pragmatic function: it promotes economy of thought and language, bundling together a set of regulative rules that we use to coordinate behavior in a set of related games. (p. 65)

In short, Guala does not advocate that institutional ideas and terms are frivolous, simply that they boil down to basic facts. He does not even claim that this formulation applies to all sorts of institutions at every level of complexity. But he does try to demonstrate that not all sorts of institutions require constitutive rules and makes an important point: social coordination seems to be the precursor to institutional phenomena and can be expressed as correlated equilibria within different coordination games expanded with correlational devices. His most important criticism of Searle's account is that his institutions can be described as rule-based, falling short of giving real incentives for why individuals follow these rules. This is in part why Guala advocates for a theory that unifies institutions-as-rules and institutions-as-equilibria approaches.

While at first glance it seems that Guala makes a solid case for why institutional phenomena can successfully be described through the merger of equilibria and rules, Searle

(2015) claims his account ultimately fails to answer several questions. I want to focus on one specifically:

> 1. All institutional facts without exception create deontic powers: rights, duties, obligations, authorizations, permissions, etc. What exactly are these, and how do they function? This is the key feature of human institutions. One of the main differences between humans and other social animals is that humans can create and act on obligations and other deontic powers. Other social animals as far as we know cannot. How does it work? (p. 508)

The first question is important because in Guala's account there are no deontic powers as defined by Searle – for Searle, deontology is what puts the behavior-modifying power behind status functions. The importance of deontology can once again be traced back to the original challenges in social ontology outlined by Przeworski: if institutions do not causally influence the agents that create and participate in them, they are epiphenomenal. If Guala explains away deontic powers by claiming that agents arrive at equilibria solutions to coordination games through individual decisions, then institutions serve no real purpose and the process of creating correlational devices remains a mystery. Mental representations are used in Guala explain the process of social coordination and individual decision-making, but do not address why social reality matters.

<u>II.10. Mental Representations as the Division Between Animal and Human Equilibria</u>

Regulative rules and equilibria don't truly explain what sets human institutions apart from animal ones. Why do humans have complex institutions and animals do not, if the mechanism is not uniquely present in humans? According to Guala, mental representations are what are unique to his account of correlated equilibria and ultimately what allows humans to create institutional reality. While animals depend on coupled stimulus and behavior – e.g. wolves and dogs react to urine as territory markers, but do so without abstract representations and agreements – human institutions can be based on any number of correlational devices. This is accomplished through representation, where representations become an intermediary between stimulus and response: "But unlike 'pure' equilibrium-based theories, this account brings at center stage the representation of the equilibrium by means of symbolic markers (rules). This way, we obtain a satisfactory, consistent, and empirically adequate conception of institutions" (Guala 2016, p. 55). Searle (2015) says Guala depends on representations

> to eliminate possible counterexamples from animals that have game-theoretic solutions to coordination problems but do not have institutions. But that is not good enough. They need to provide a theory of representations that will explain their function in institutional ontology. Are representations institutional facts? And what sort of an institution is the institution of representing? (p. 513)

The crucial element of human institutions that must be understood is what allows representations to propel humans from non-institutional, non-cooperative behavior to cooperation and institutionality. This step is what Searle is also interested in, which he argues creates institutional deontology (Searle 2010; Searle 2015). An adequate account of institutions has to acknowledge the role of language in creating the rules that govern institutions, and how language itself came to exist as an institution. Guala's examples of equilibria and rules all involve preexisting institutional concepts that are left unexplained. For instance, organization into the Nuer and Dinka tribes already requires a complex institutional ontology, recognizing distinct groups with names. In addition, in the description of his scenario about the development of the institution of personal property involving the Nuer and Dinka, he describes the cattle belonging to each tribe; this implies that personal property already exists in the form of owning animals. Hence the entire scenario is not a hypothetical "pre-institutional" situation, but a set of circumstances already steeped in human institutions, specifically with the one whose development its attempting to explain. Hence, Guala relies on already existent institutions (presumably built on representations themselves) to explain further institutions. The logic is circular.

II.11. Conclusions for Searle and Guala's Social Ontologies

As discussed in this section, the red thread running through both Searle and Guala's work is fundamentally anchored in their use and understanding of mental representation. Despite having two distinct approaches for building their respective social ontologies, Searle and Guala both depend on traditional representation. This problem points to a larger issue that, in fact, both Guala and Searle have to contend with. By subscribing to an traditional account of representation – and thus to internalism – they are forced to (re)connect and also separate the individual to the social, the mental to the physical, and the human to non-human. This results in either overly top-heavy institutions like in Searle, or overly bottom-up, mechanistic institutions like in Guala.

Guala devotes most of his account to explaining how institutions persist and allow for cooperation by appealing to equilibria. It is important to point out this aspect of social ontology, but without a more robust explanation for how stable representations are created, his account remains incomplete. For Searle, the creation of representations is explained through appeals to intentionality (individual and collective), and is very much at the center of his ontology: “all of human institutional reality is created and maintained in existence by (representations that have the same logical form as) [Status Function] Declarations, including the cases that are not speech acts in the explicit form of Declarations” (Searle 2010, p. 13). The persistence of such representations remains mysterious; as Guala points out, an essential aspect of institutions is to allow for persistent cooperation to occur, which has to have incentives for all individuals involved. If we are to understand social ontology, we must take a deeper look at what is so unique about how we represent them cognitively, or if the cognition of representations must be rethought altogether.

III. Representations

III.1. Introduction

Since both Searle and Guala rely on representations to formulate their accounts of institutions, it is worth examining this concept more closely. There is an extensive literature about the nature of mental representations, so I will be examining a small sample of papers that are relevant to the discussion of representations and their role in social ontology. Namely, the debates between representationalists and anti-representationalists along with Bickhard and Terveen’s (1995) book on encodingism and representation will provide the background for current views on and approaches to representation in social ontology and cognitive science in general. Since the main argument of this paper is to examine representations in social ontology through the lens of the Free Energy Principle, this section will thus focus on establishing the reason for why the FEP account offers a strong alternative characterization of representations that could potentially answer challenges posed by anti-representationalists and stand up to critiques of encodingism, while continuing to employ a theoretically useful form of mental representations.

III.2. Criticism of Encodingism

One angle of criticism aimed at the representational model for social cognition comes from challenges to a representation-based conception of theory of mind. More precisely, in their book, *Foundational Issues in Artificial Intelligence and Cognitive Science* (1995), Mark Bickhard and Loren Terveen make program-level criticisms aimed at the entirety of cognitive science research: they claim the field takes encodingist representation to be the basic building block of cognition which, they argue, cannot accurately reflect our cognitive mechanisms. Bickhard and Terveen examine the structure of representation in cognitive science and artificial intelligence, claiming that it is logically recursive.

The authors' description of the most commonly used version of representations is one which can be described as encoded symbols or encodings. The encoded symbol represents something in the world (an event, a color, a concept, etc.) and thus carries representational content, corresponding to said phenomenon in the world. Symbols are expressed through symbol types, which are "formal 'shapes' whose instances can be physically distinguished from each other within whatever physical medium is taken to constitute the material system" (p. 11). As an example, they give the morse code "• • •" which corresponds to the letter "S", thus making it a representation of the character S. In cognitive science research, the authors claim, this representational structure is taken and applied to the mind: states of the mind of an agent are encodings of some phenomenon in the states of the outside world (an event, a color, a person, etc.). Everywhere in cognitive science research where representations are present, implicit assumptions of their encodingist nature come with them.

This is still one of the most dominant frameworks in cognitive science research, as it fits into the Computational Theory of Mind (Fodor 1975; Pylyshyn 1984), and posits that the brain functions as a kind of computer, manipulating an internal set of symbols (representational states) through a combinatorial syntax and semantics, constituting a language (Dołęga, Roelofs & Schlicht 2018; Williams 2018). This version of representationalism implies that the job of cognition is of replication – creating internal representational states to be manipulated to best reflect the outside world. It also places cognition between action and perception – we perceive outside stimuli and turn it into internal representations, perform the appropriate computations and produce some action as the output. Furthermore, the brain in this case is the seat of cognition, with all stimuli going through the sensorium toward the brain. The brain, thus, does not have any impact on the sensorium, which simply reports, as accurately as possible, the outside world. The brain's purpose is to construct representational states from the incoming sensory information and manipulate them in order to then produce the proper behavior.

The authors take issue with this for the reasons stated earlier – that it is unclear how representations can carry representational content and which thus makes it difficult to imagine how representations in the mind can causally influence the agent. They add that the encodingist conception is logically flawed in itself:

> Encodings can carry representational contents, and already established encodings can provide representational contents for the formation of some other encoding, but there is no way within encodingism per se for those representational contents to ever arise in the first place. There is no account, and — we argue — no account possible, of the emergence of representation. (p. 13-14)

One encoding can stand in for the encoded content of another encoding, but at some point, there has to be a bottom level of encoding. It is this bottom level that encodingism cannot provide an explanation for:

> Consider some element X of such a purported logically independent, bottom level, foundation of encodings. On the one hand, X cannot be provided with representational content by any other representation, or else, contrary to assumption, it will not be logically independent — it will simply be another layer of stand-in encoding. On the other hand, X cannot provide its own content. To assume that it could yields "X represents whatever it is that X represents" or "X stands-in for X" as the provider and carrier relationship between X and itself. This does not succeed in providing X with any representational content at all, thus does not succeed in making X an encoding at all, and thus constitutes a logical incoherence in the assumption of a foundational encoding. (p. 14)

Without a way to provide the bottom-level of representation with content that does not come in the form of another representation, encodingist representations are undermined by a circular logic.

For this reason, the theoretical commitment of Searle and Guala's accounts to mental representations as fundamental to social cognition becomes precarious. The reason the issues discussed in the previous section arise in both Searle and Guala are because of this theoretical commitment to traditional representation, which exhibits many of the issues of encodingism and seemingly cannot supply a proper explanation for its content.

III.3. Arguments for Encodingism

A rebuttal to this argument would be to consider representations as encodings that simply "stand-in for things in the world" (p. 15). The problem with this is that encoded representations can only stand in for other encoded representations (as in the case of "• • •" standing in for "S"). They cannot stand in for things in the world, because things in the world are not representations. A mental representation of something is not structurally equivalent to the thing that it represents – you cannot use a representation of XYZ for the same things as XYZ itself. So, encodingism only addresses "issues of manipulation and transformation of already constituted carriers of representational content — carriers for some interpretive, intentional agent" (p. 17). We are left without any real explanation of mental representations' nature, function and emergence. What Bickhard and Terveen argue is that instead, there needs to be an account of a provider and interpreter of representational content: how the content of mental representations is provided, how it is interpreted and how it functions in cognition.

So, what impact does the use of encodingism have for cognition research? Bickhard and Terveen (1995) claim there are several logical consequences of using encodingist representations, the first one being that "to check their accuracy would require that we have some epistemic access to the world that is being represented against which we can then compare our encodings, but, by the encodingism assumption, the only epistemic access to the world that we have is through those encodings themselves" (p. 19-20). Without an explanation for the emergence of representations from non-representational reality, we get no way of knowing the validity of an encodingist representation. One possible counterargument to this would be that we don't, in fact, have any epistemic access to our encodings, and thus live in a reality constructed entirely out of our own encodings. This, however, requires an extremely solipsistic view of the world. If we don't have independent epistemic access to the world, we have no idea how to construct a representation. One version of this argument is that we are not able to construct copies of the world without knowing what the world is (Piaget 1970). Lastly, with the encodingist framework, emergence becomes impossible – this is because, as stated earlier, if a representation is an encoded symbol, it can only be based on another encoded representation.

Bickhard and Terveen (1995) claim that there is no questioning the fact that representations exist, so there must be a way to explain how they emerge from non-representational reality.

There have been attempts to address the incoherence of encodingist representation, and the authors outline several – and explain why they don't work. First, Fodor's (Fodor 1981; 1983) theory of innate representations tries to address the encodingist incoherence by shifting the burden of representational emergence to evolution. Basically, the content-provider of mental representations becomes evolutionarily evolved – innate encoding generators – which cannot be learned. These innate encoding generators are the basis of all other encodings. Also, the "combinatoric space defined by [this] set of basic encoding generators constitutes the entire possible representational world of an epistemic system" (Bickhard & Terveen 1995, p. 25). However, Bickhard and Terveen argue that this solution misses the point of the encodingist incoherence, which is logical in nature. Therefore, leaving the heavy lifting of solving how the bottom level of encodingist representations can emerge to evolutionary processes does not actually solve the incoherence. It does not answer the fundamental problem of addressing the emergence and function of representations.

Another attempt from Fodor to explain encodingist representation is to frame it in terms of what is being represented (Fodor 1981). Our knowledge of what is represented is based on our knowledge of the world, or, of our knowledge of chemistry and physics. Our representations aren't properly defined until we gain knowledge of the world, but in order to gain knowledge of the world, we need to represent it somehow – the encondingist contradiction still stands. Bickhard and Terveen (1995) then move on to the next purported solution – direct reference. This explanation views representations as direct encodings of things in the world, with the encodings being some sort of true, basic names that take away the burden of content for the representation – there is no content. It is unclear how these sorts of representations come into being or how they function. As such, the idea of direct reference did not gain much traction as a framework for representations in cognitive science and was thus left underdeveloped.

There are a series of conceptualizations that are more popular, revolving around the "observer-user". These conceptualizations attempt to construct representations in relation to an "observer-user" in different ways. External observer semantics posits that there are no real content-bearing representations inside the observer, instead, “the observer-user knows that certain of the inputs, and certain of the outputs, are in such-and-such a correspondence with certain things in the world, and are thus available to be taken by that observer-user as encodings of those things in the world” (p. 27). The issue is that not having representational content within the observer-user “makes phenomena such as the generation of representational content, and intensional [sic] stances with regard to representational content, impossible to even address” (p. 28). Internal observer semantics, on the other hand, adds certain internal elements to the system (the observer-user, its environment, and the body), which are then manipulated and transformed by the system. The representations in this case can be about correspondences not just between inputs and outputs, but between elements within the system itself. The issue is that these representations are still only accessible to the observer-user, but not to the system itself. Once again, we are stuck without an explanation for how the representational content is constituted in a way that can function to have causal influence on the system.

Generally, the attempts to justify the encodingist version of representation suffer from, Bickhard and Terveen (1995) argue, a confusion around epistemic versus factual or causal correspondences. Transduction aims to base encodingist representation in the transformation of energy in the environment by “system transducers” – such as sensory receptors – into internal encodings. This is problematic, because “there is no explanation or explication of how the ‘elements that are in correspondence’ – e.g., products of transductions – could constitute encodings *for the system*, not just for the observer-user” (p. 32). So, there is no account for how the transduction turns factual/causal correspondence (transforming energy from the outside world) into epistemic correspondence (knowledge usable for the system in its functions, i.e. emergent representations with content). In general, for the encodingist account of representation, there seems to be an issue connecting the epistemic agent with the physical, causal agent.

III.4. Anti-Representationalism

In response to the thorny issue of understanding how the epistemic (representations) can reasonably inform the cognitive system, some philosophers have opted to argue against the very idea of representations. This angle of criticism comes from 4EA approaches[3] which question the way brain, body, and environment affect each other. More specifically, they expand cognition beyond the confines of the brain to encompass this interaction between brain, body, and environment: cognition is extended in a way so that its locus is not simply in the brain, but spread out over the agent and its environment, both being equally causally important in constituting it. Furthermore, enactivism breaks up the linear chain of sensation, representation, and action, advocating instead for a more fluid interaction that involves the interaction between brain, body, and environment. This interaction is not linear because bodily perception might affect processes in the brain and vice versa: action might influence the environment and the environment perception. More radically, some enactivists argue that perception, action, and cognition are accomplished by the whole organism in tandem with the environment, without the need for representations (Varela, Thompson, & Rosch 1991; Noë 2004; Hutto & Myin 2013; Hutto 2005).

Hutto, Kirchhoff, and Myin (2014) advocate for the acceptance of REC (Radically Enactive Cognition), asserting that doing so offers a valuable opportunity to redefine the extended mind dispute in the field of philosophy of mind. In their seventh chapter, they argue that embracing REC serves as both a motivation and an effective method to move beyond the ongoing "extended mind" debate. Their book identifies a crucial hindrance to progress in this debate, which is the persistent commitment to the concept of cognition being necessarily content involving (CIC). This notion has been embraced by both sides of the debate (Hutto and Myin 2013).

---

[3] Advocates of 4EA cognition challenge the traditional view that cognition is an isolated, abstract process occurring solely within a central unit in the brain, which is usually associated with a kind of Cartesian framework. This conventional view is typically associated with functionalism, which holds that cognitive processes are entirely defined by their functional roles and can be studied independently as a distinct level of analysis. In contrast, proponents of 4EA cognition argue that the cognitive processes examined by contemporary cognitive science are fundamentally shaped by an agent's bodily form, biological and physiological characteristics, the structure of their surrounding environment (whether natural, technological, or social), and their active, embodied engagement with that environment (Newen, De Bruin, & Gallagher 2018). They also advocate that these processes to all have an essentially experiential aspect to them, hence the A which stands for affective.

According to the REC analysis, relinquishing the assumption of CIC leads to a radical shift in our comprehension of cognition's scope, demanding a complete reimagining of the standard extended mind discourse. Such a change holds significant implications for proponents of the extended mind theory and its internalist opponents. The core reasoning is straightforward: If we consider that basic cognition is not representational in a contentful manner, then the main foundation for internalism is undermined. Internalists lose their strongest rationale for asserting that cognition, particularly in its primary evolutionary and developmental forms, is an "inner" process, implying cerebral localization of content vehicles.

While some advocates of the extended mind theory may also hold the belief that cognition involves content, they differ in their views on the location and extent of such content vehicles. From an REC perspective, the distinction between vehicle and content does not apply at the level of basic minds; where content is absent, so are the vehicles of content. Therefore, if it is discovered that basic forms of cognition lack content, the extended mind debate would undergo a transformation.

Thus, enactivism rethinks sensorimotor capacities – how perception and action unfold in a coupling with the environment to produce complex action possibilities for the organism. In other words, perception is already structured according to highly specialized realizable actions relevant to the organism instead of simply mirroring reality (Hutto, Kirchhoff and Myin 2014). As for action, it is actively a part of perception (eye saccades, head movements, moving the body to touch something) and not simply the output of cognition. In this way, it challenges the position occupied by mental representations, which is currently characterized as intermediary, with representations used in the feed-forward chain of perception, cognition, and action, existing exclusively in the brain. Cognition, for enactivists, is instead realized through an interaction between brain, body and environment. Enactivists go further and pose the question: if cognition is spread out between brain, body, and environment, just how constructive is it to think of cognition in terms of internal mental representations? The radical enactivist view cites a series of problems with current conceptualization of representation-based cognition. Through a series of challenges, the enactivist, anti-representationalists question the necessity for mental representations.

Williams (2018) outlines three basic challenges posed by anti-representationalist, related to representational function, representational content, and cognitive function. The first challenge about representational function has to do with what a representation accomplishes that cannot otherwise be realized through the physical processes and structures implicated in cognition. The first part of this challenge asks what sort of physical structures can be characterized as genuinely representational. As Williams puts it,

> if internal representations are to genuinely *explain* intelligent behaviour, their *effects* on behaviour must be a function of their representational *status* as content-bearers. Without this, a *representational* explanation of the system's behaviour would be causally redundant. (p. 143)

In other words, if parts of a cognitive system act representationally, why is their representational status causally important in generating behavior? A second element to this challenge is about the neat, triadic relationship between the representational vehicle, its target, and the cognitive system that interprets the vehicle in order to act on its target. If the representational vehicle's representational status is exploited by the cognitive system of which it is a part, how can the system make use of it (O'Brien & Opie 2015; Williams 2018)? Historically, this challenge was phrased as the cognitive system needing inner homunculi as intelligent as the system they are a part of (Wittgenstein 1953). These challenges are then the grounds for the claim that the structures implicated in representation in contemporary cognitive science are not actually filling a representational role, so it is thus not necessary to think of cognition in representational terms (Ramsey 2007).

The second challenge is about how the content of representations cannot be understood in a naturalistic metaphysics, leading anti-representationalists to declare that they thus do not exist. The challenge seeks "to identify the natural properties, relations and processes that determine the intentional properties of internal representations without circularity?" (Williams 2018 p. 145) While Fodor (1975, 1987) and others attempt to address this question through "naturalistic psychosemantics"[4], anti-representationalists remain unconvinced. The main concern is how to

---

[4] In *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (1987), Jerry Fodor advocates for a conception of mind in which certain basic beliefs/desires are based on representations and acquire content mainly

map symbols existing inside brains to things in the environment, with the main issues being determinacy, shared contents, and the possibility of misrepresentation (Williams 2018).

The third challenge addresses cognitive function, with the main contention being that the function of cognition is fundamentally not to copy the world through representations, but to act on it. Thus, it is unnecessary to implicate representations in the process. Engel et al. (2016), suggest that cognitive science is going through a "*pragmatic turn* away from the traditional representation-centered framework" toward one that instead views cognition "as subserving action". (p. 1) More specifically, what is meant by action is regulation and control (Williams 2018; Anderson 2014; Anderson 2016; Cisek 1999). Cisek (1999) outlines several contentions with the traditional, computational model, but also brings up the issues of meaning and consciousness (among others) that remain impossible to contextualize in the input-cognition-output model.

III.5. Regulation as an Alternative to Encodingist Representation

Instead of the computational model that is part and parcel of encodingism, self-regulation and self-creation can shift the emphasis of cognition toward action-oriented self-maintenance. Cisek proposes to think of cognition in terms of homeostatic regulation within a certain environment: living systems (like humans) attempt to regulate their internal states (e.g. pH, nutrient concentration, temperature, etc.) to keep them within a range necessary for their existence. This regulation requires energy and active engagement with the environment. Homeostasis can thus involve interaction with the environment in order to seek out necessary resources for the regulation of internal states. For humans, such regulation obviously goes beyond simple chemical indicators such as pH, leading us to consider more long-term goals and to display complex behaviors, but the basic idea of cognition serving action/regulation remains true. I will return to why rethinking cognition in terms of regulation is useful and how it fits as an

because the content accurately reflects the state of things in the world. He uses this to explain how meaning is created in the mind.

explanation for complex human behaviors in the section IV, but for now the main takeaway is the possibility of this reformulation.

Cisek (1999) explains succinctly what lies at the heart of the challenges to the traditional model: addressing the mind-body problem. He argues that cognitive science has roots in the 17th century philosophical foundations of psychology, which advocated for the separation of mind and body. The belief in this split defined psychology, with psychologists even today working "within scientific traditions which grew up in the context of [a soul]" (p. 2). This has had implications for discussing and studying behavior:

> The assumption of a non-physical *Mind* compelled philosophers to conceive of two interfaces between it and the world: *Perception* – which presented the world to the mind; and *Action* – which played out the wishes of the mind upon the world. […] When dualism was eventually rejected, the concept of the non-physical Mind was replaced with a mechanistic concept of Cognition, but the shape of the architecture remained, leading to the Perception-Cognition-Action model of behavior (sometimes called the "sense-think-act" model). (p. 2-3)

Coming from the cognitive revolution, with advocates like Fodor and Pylyshyn, this "mechanistic" conception of cognition introduced the idea of computation being a solution to the mind-body split that had veered psychology into extremes (introspectionism and behaviorism)[5]. This computational model provided an answer for many concerns in psychology:

> First, the computer metaphor provided a candidate mechanism for how Cognition operates – it operates like a digital computer program, by manipulating internal representations according to some set of rules. Second, it allowed discussion of non-dualist internal states – e.g. memories and temporary variables. Third, it provided an inspiring metaphor for functionalism – mental entities are like software while physical mechanisms are like hardware. This same metaphor also

---

[5]Introspectionism took shape during the 19th century, with the development of quantitative introspective methods, especially concerning sensory consciousness, that the study of the mind began to evolve into a systematic, laboratory-based science (Schwitzgebel 2023). Pioneers like Helmholtz, Fechner, and Wundt sought to understand the complex relationships between stimulus and sensation by studying experience. Behaviorism insisted instead that any psychological hypothesis must be supported by observable behavior. For them, there is no meaningful distinction between different mental states (such as beliefs or desires) unless those states are linked to distinguishable behavioral differences. Costall (2006) argues that the story of introspectionism-behaviorism-cognitivism is not quite right, as introspectionism did not dominate 19th and early 20th century debates in psychology, nor was it completely detached from behavior. As such, behaviorism was not an answer to the hegemony of introspectionism. In the same vein, the approaches emerging out of the cognitive revolution not so much as supplanted, but extended behaviorism. But for the argument of the thesis, these historical movements are not paramount.

> provided a quick way to bridge the gap between psychology and biology – psychological phenomena are the software running upon the biological hardware. Finally, it provided a mathematical formalism that gave psychology some long-desired rigor – the language of predicate logic and information theory. (p. 4)

The computational model addressed these concerns well, but as stated earlier, could not make any space for meaning or consciousness. Without addressing how these features of human cognition fit into the picture, the computational model of mental representations remains shaky ground to build theories of social cognition on. As Hutto, Kirchhoff and Myin (2014) put it:

> This is what justifies thinking of minds as naturally extensive—the idea that cognitive activity always already entangles embodiment, action, and world-involving resources and does not restrict itself only to what is inside the individual organism. Does this mean the extent of cognition has no limits? Of course not. And those limits can be revealed by empirical experiment. Even if one were able to see for miles one might not be able to see for leagues. There will be empirically discoverable species-wide and individual differences with respect to the limits of cognition—limits that can and should (and are being) actively empirically explored. (p. 10)

Conceptually shifting from internal representations to extended – or, as some radical enactivists prefer, extensive – minds, that involve complex causal interplay between brain, body, and environment is not simply a theoretical one, but an ontological shift that has empirical consequences.

The role of representations in cognition and their ontological significance continues to be debated. Contributing to the debate on the necessity of representations is beyond the scope of this thesis. The analysis of these writings serves to highlight that whether or not representations are a basic necessity to cognition, how we conceptualize them has reverberations for the ontological and empirical foundations for theories built upon them. This has especially important implications for social ontology because it examines extremely complex, multi-faceted forms of cognition involving multiple agents and a massive variety of environmental factors. It is tempting to create explanations of institutions by reducing their complexity into elements that can be empirically tested – individuals, their beliefs/desires/intentions (their mental representations), or their choices and their ideas about others' choices (again, mental representations) – but if it comes at the cost of eroding the theoretical and ontological viability of these explanations, it is worth to consider reevaluating their theoretical building blocks.

III.6. The Relevance of Anti-Representationalism for Searle and Guala

The challenges posed by interactionists and enactivists are important for Searle and Guala not only because they endorse a theoretical stance that treats representations as stable, internal and contentful, they both implicitly use a feed-forward conception of representations that rely on the computational model, where sensory inputs are turned into action outputs. Whether these representations are social in nature does not change the underlying problems. While Searle and Guala do not spend much time diving into the representational framework they endorse, they nonetheless indirectly grapple with the challenges brought up by Williams, Cisek, and the proponents of enactivism and interactionism. What sorts of structures are responsible for making shared representations that influence people's behavior? How exactly do the representations – more generally, their content – in Guala and Searle's models function to drive behavior? How do misrepresentations work in models that depend on representations being stable, truthful reflections of reality?

The representations in Guala and Searle's theories do not provide rich explications for the cognition involved in social ontology; Searle makes an attempt to flesh out his concept of representation, which is based on research about shared intentionality. But Searle runs with the basic idea of shared intentions, and constructs a rule-based account of institutions, where agents represent such institutions through constitutive rules. There is a considerable gap between the concept of shared intentionality and a system of institutions based on stable constitutive rules. Having different, uniquely human abilities that fall under the umbrella term of shared intentionality does not necessitate a system of constitutive rules and social facts. At no point does his account clearly address the nature of the representations he imagines, and as Pacherie (2005) points out, it is unclear even in his account of collective intentionality how a we-intention vs. an I-intention really functions within an individual and between two individuals.

Searle appeals to the special, socially constructive properties of we-intentions, yet maintains that they exist within an internalist representational framework, where individual minds are the building blocks of institutional reality. For Searle, "constitutive rules" are the natural consequence of turning we-intentions into institutions. What's more, Searlean institutions carry deontic power: their collective nature brings about their ability to motivate individuals to act in desire-independent ways (Searle 2010, p. 9). This is what allows people to construct a world in terms of social facts. However, we are left to wonder exactly how the complex intersections of collective intentionality, constitutive rules and social facts are realized through cognition.

It seems Searlian we-intentions (representations with specific content regarding the agent's role in institutional reality) work in a way where they allow in the joint construction of a social reality. This is where we encounter issues with representational content, function, and cognitive function. Are we-intentions different for individual minds or are they the same (is their content the same)? If they're different, what allows them to create this new, shared reality? We see the struggle to have representational content, while also maintaining that this content (which is supposedly shared) somehow leads to unique institutional rules. How can these rules be "arbitrary" and have deontic power, as Searle claims, if they are created as a result of shared representational content? If they're based on shared representational content – which, according to Searle, is reducible to brain chemistry – then basic facts should determine the course of institutional facts, not vice versa. Thus, the heavy lifting in this theoretical framework is being done by these special, we-intention representations, which somehow try to make sense of the paradox of having representational content and top-down normativity.

The key takeaway here is not that humans cannot engage in joint intentionality, but that this process cannot be distilled into stable, internal representations whose contents are we-intentions that allow for socially constructed reality. Joint intentionality has to happen between two (or more) individuals, in an interaction that exists spread across the agents and the environment. Deontic powers cannot just appear due to the shared, collective nature of we-intentions because their internalist, representation-based conception does not allow for the construction of a coherent account of social ontology.

Guala's theoretical approach seems very different at first glance, as it appeals to incentives and individual decisions added together to form a functional approach to the basis for a social ontology. Interestingly, he claims that his equilibrium account "does not even presuppose that the players can form representations (beliefs) of other people's preferences, expectations, and actions. This is particularly handy when we use games to model the interactions of nonhuman players, such as animals or organizations (e.g., firms)" (Guala 2016, p. 24). However, as discussed at the beginning of this chapter, he also adds that representations are crucial in allowing humans "to decouple stimulus and behavior" through representing rules, setting them apart from animal behaviors that can also form equilibria, earning his account the title "rules-in-equilibria". To sum it up in Guala's own words:

> The rules are symbolic markers that represent equilibria (or parts of equilibria) and help the players use a particular coordination device. Unlike in rules-based theories, the concept of pattern (equilibrium) is central in this theory. But unlike "pure" equilibrium-based theories, this account brings at center stage the representation of the equilibrium by means of symbolic markers (rules). This way, we obtain a satisfactory, consistent, and empirically adequate conception of institutions. (p. 55)

Essentially, Guala is describing internal encodings – "symbolic markers", or rules – as the crux of his account, with these symbolic markers acting as the unifying force between pure equilibrium approaches and pure rules-based approaches. As Bickhard and Terveen's (1995) analysis and critique of encodingism shows, however, it is shaky ground to build upon, as there is no way for representational content to arise within an encodingist framework. It is difficult to build an account of social ontology without a clear explanation for how the supposed representations that are the backbone of such an account come to be.

Whether on purpose or not, Guala's account uses a traditional concept of representation, which behave like encodings. This is because he does not really expound on what he means exactly by representations or how agents' cognition realizes the sorts of representations needed for a game theoretic conceptualization of institutional ontology. He discusses how institutions can be defined, contrasting folk theoretic notions with expert identifications of core empirical properties, which are often at odds with each other. (Guala 2016). He refers to Sally Haslanger's (2012) work to make sense of the friction between empirical and common sense understanding of an institution. Haslanger explains that for an analysis of social institutions, there are three key elements to consider: the "operative concept," the "manifest concept" or theory, and the "normative concept." The operative concept refers to the actual practices linked with the term's use. The manifest concept represents the theory or stereotype individuals use to understand the term. Lastly, the normative concept is the ideal version of an institutional entity we aspire for. These concepts often share commonalities, but inconsistencies are not uncommon. Guala (2016) adds that, on Haslanger's view: "One can be a social constructionist, because gender and race are socially maintained, rather than biologically determined, categories; and one can be a realist, because the terms 'race' and 'gender' refer to real social kinds, that is, classes of entities that share important properties and whose commonalities are sustained by clearly identifiable social mechanisms" (p. 189-190). There is indeed something paradoxical about this view, since constructionism and realism seem diametrically opposed to each other. It is this paradox that Guala hopes to resolve with his equilibrium approach.

Guala highlights just how difficult it is to try and define a social institution in seemingly empirical terms, only to find out that in reality, people think of it differently and, at some point, the institution may take a different shape altogether. Universities, for instance, could be described 100 years ago as nearly exclusive storehouses of academic knowledge, and the centers of scientific research. It could be argued then, that universities were places with libraries to store books which contained knowledge, classrooms and dormitories for students to learn and live in, and offices and laboratories for academics to conduct their research in. What may have been true and a good working idea of a university then, seems to have lost its meaning today – universities are no longer exclusive depositories of knowledge, as books, journals, films, audio sources, etc. have migrated from a physical to a digital medium and are widely available on the internet for anyone. Students no longer have to physically live or learn at their universities, as online courses make it possible to attain a complete university degree more-or-less wholly from the comfort of one's own room. Academics and researchers now do not necessarily need physical spaces to conduct research, as online surveys and programs have made it possible to recruit participants from all over the world. And for those pursuing research requiring physical equipment, universities are no longer the exclusive powerhouses of innovation – many companies have grown to such a size that they can and do have their own labs, and their own research departments full of scientists.

It could be argued that my attempt at characterizing a university simply missed the essence of what a university really is, but the difficulty in trying to find the objective parameters of any institution is a challenge, to say the least. Anthropologists, for instance, have struggled for decades to successfully define marriage:

> Over the years as well, anthropologists have debated what lies at the very core of marriage as an institution and as a relationship, whether it is economic sharing, sexual access, childbearing and parenthood, or reproductive success. They also have debated whether marriage is best described in terms of sentiment and subjectivity, social or mental structures, or transactions and strategic choice. (Levine 2014)

Guala's attempt at trying to integrate constructivist and objectivist approaches is to suggest that social institutions come about/are created as an answer to a problem of cooperation, where there are adequate incentives for all agents involved for an equilibrium (or equilibria) to emerge and become stable.

While seemingly a fresh perspective, Guala is essentially describing a functionalist approach to institutions – they solve the problem of coordination. Coordination is the basic problem that must be solved by institutions, thus they serve to maintain or enable such coordination. That is their function. This is not to say he had nothing new to add – he has fleshed out his functionalism with mechanistic explanations of rules, game theory, Schelling focal points and Lewis' ideas on convention. This way, he asserts that institutional function arises simply from mechanistic interaction.

This is problematic because function cannot just emerge from mechanistic interaction, as anti-representational arguments highlight. Whether on the level of mental representations or social institutions, the idea that function cannot arise from content holds. Additionally, Guala is advocating for an approach – a functionalist account of social institutions – which has been tested around a century ago and later largely abandoned by anthropologists and sociologists. Why? Anthropologist I.C. Jarvie (1986) outlines a critique of functionalism in anthropology, which has bearing not just on ethnographic research, but also on why glossing over the co-determination between mechanism and function – as Guala does – is questionable. Jarvie explains that at its core, a functionalist approach to institutions is problematic for two reasons. First, it is, in its essence, a deductive form of investigation, which Jarvie argues, not only cannot provide satisfactory (non-circular) reasoning but cannot even be considered logically valid. A functional explanation for a social institution (going to church, for instance), could go something like this:

> (x) People in all societies must express and reinforce their social solidarity TRUE
>
> (y') Going to church is one possible way of expressing social solidarity TRUE
>
> (c) People go to church FALSE or TRUE
>
> Whether (c) is true or false, it does not clash with (x) and (y'); therefore the derivation is not valid because a counterexample, an inference of the same form with true premises and false conclusion, can be constructed, since (y') covers only some possible cases. Thus if church-going is only one possible way of expressing social solidarity, it does not follow that church-going is explained by the need to express social solidarity. (p. 132)

In and of itself, a functional explanation of a social institution does not seem to have logical validity. Such an explanation can of course be amended and refined, but it ultimately has to reduce a complex institution into a set of observer-independent motivations and behaviors.

Jarvie further argues that a functionalist explanation does not have any explanatory power because it does not provide new understanding of the facts it attempts to explain:

> A theory makes a satisfactory explanation when it is not circular, not ad hoc, and is testable independently of the facts it is intended to explain. But a functional theory does not go beyond the facts it intends to explain; therefore a certain tinge of circularity, or of ad hoc-ness, is an inevitable consequence. To be told that the function of church-going is to express and reinforce social solidarity and that the main test of the desire to express and reinforce social solidarity is church-going is to get into a circle which cannot be broken in favor of a "deeper" explanation. Circular, or ad hoc, or nonindependently testable explanations do not tell us anything new; they are methodologically unsatisfactory. (p. 134)

Indeed, this circularity is present in Guala as well, if we examine his account of how streetlights function as an institution. Drivers need to coordinate intersections, so streetlights are added to tell them who goes and who stops. So, if the function of streetlights is to tell drivers when to stop and go, then must drivers' need to decide when to stop or go be necessarily in the form of following the colors of a streetlight? Of course not – stop signs, roundabouts, "priority to the right" rule, etc. – could fulfill the function of coordinating drivers' sequence in crossing an intersection. So how do we arrive at this specific solution?

Guala’s attempt to explain why we settle on certain forms of cooperation over others is to appeal to the work of Thomas Schelling (1960), which deals with the concept of focal points in game theory. As outlined earlier, Guala’s basic argument is that Schelling focal points in the environment provide agents with the foundation for shared solutions to coordination. This explanatory leap does require quite a bit of generosity, as even with a coordination problem as simple as drivers’ crossing an intersection, the existing solutions are diverse and most definitely not exhaustive of all the possible solutions to this particular coordination problem. This is somewhat of a theoretical sleight of hand, as the jump from a notion of shared focal points to real, specific solutions to coordination problems in complex institutions is quite big. In fact, this issue is what brings us full circle: the underlying idea of shared focal points being the basic building blocks of coordination problem solutions does very little to explain the variety of unique solutions that people invent to deal with coordination problems. Guala is, in a sense, turning something in the environment (a Schelling focal point) into a representation that is universal – all of us see a certain thing the same way and represent it in the same way so that we can then build conventions on said representation. This precisely describes a feed-forward (environment-to-mind), internalist, and encodingist sort of representation that has been critiqued by proponents of enactivism and interactionism.

Thus, functionalists such as Guala, who fall into the rational action neoinstitutional category, could be accused of an overly static view of institutions. As outlined by Akiyama and Kaneko (2000):

> Because the games themselves in this case are static in nature, it could even be said that one of the main reasons for the success of these studies is the “static” representation of the games. [...] Certainly, if we think of social phenomena in the real world, there are many examples that can be investigated very effectively within the framework provided by static game representation. However, on the other hand, examples of phenomena that cannot be completely analyzed in this framework also exist. This is because the game environment of the real world that we actually live in is not fixed, and changes with every action that is carried within a strategy or any change in the value judgment system used by a player. In real-life games, the mutual interactions between the game environment and the players can have a large effect. (p. 223)

If we take an encodingist, internal framework for representation, the “universal” points of commonality in the environment imply that different agents’ representations of them will also be the same. If this is the case, then a computational view of the mind, where representations are manipulated in order to produce action, would in theory produce very similar, if not identical actions. How can institutions be flexible and often quite different given extremely similar circumstances? Guala explicitly endorses a stimulus-representation-action model of cognition, which he goes as far as to say is the key difference between animal "faux" institutions that are based on coupled stimulus-action, and real, flexible human ones.

The theoretical problem for building institutions out of behaviors that are the result of mental representations, however, is that it hints at the representations being stable and reflecting an objectively true reality identified in external, true properties inherent to the phenomena being represented. Such representations then somehow causally lead to action, with different agents producing similar behaviors that fall into choices within a Nash equilibrium. Trying to avoid being accused of an overly static view of institutions, Guala explains that the representation-based rules he describes can be "bundled" together to make more complex, vague notions of larger institutions, such as marriage, where formal and informal rules mingle to form an ever-changing entity (p. 67-68). This, however, does not solve the logical problems posed for encodingist representations.

This is not to say that the functionalist view does not shine a light on an extremely important aspect of human institutions: they are, in fact, used successfully by very different people, and institutional newcomers can easily learn rules and adapt their behavior to seamlessly fit into the behaviors required for fruitful social cooperation to occur. Without some universality and similarity in our cognitive capacities, it would be impossible to understand different customs or find solutions that work reasonably well for groups ranging from just two individuals to entire societies. However, it is important not to fall into the social constructionist neoinstitutionalism extreme: just because a mechanistic, functional explanation is incomplete, does not warrant the idea that institutions are then unconstrained and determined entirely by the rules set by their participants. How can we arrive at mutually agreeable rules and why is it that we then follow them?

III.7. Towards an Alternative Framework of Representation

As discussed, both Searle and Guala's work, representing Social Constructionist and Rational Action Neoinstitutionalism respectively, show a reliance on the traditional feed-forward, stable, internalist form of cognition that is built on the formation and manipulation of mental representations in order to produce actions. This form of representation does not include an explanation for how its content is structured or how it can emerge, and how this content has causal bearing on an individual's behavior. There is also no explanations for how representations function cognitively in the mind – what exactly is a representation and how does it work within the architecture of our cognition (in this case, social cognition specifically).

So how can this issue be addressed to move forward and place institutional foundations on more cognitively secure ground? As interactionists and enactivists have pointed out, in order to form an ontologically sound understanding of institutions, the interaction between individual and (social) environment must be taken in its entirety to be the locus of cognition. If the representationalist foundations of Searle's and Guala's theories cannot be sustained, what kind of framework could replace them? Such a framework must avoid encodingist assumptions, account for the dynamic interplay between individual agents and their environments, and explain how large-scale social regularities can emerge from local interactions. Recent work in cognitive science, particularly the Free Energy Principle, provides just such a model. FEP reconceives representation not as a static inner encoding, but as an ongoing process of prediction, error-correction, and adaptation. This shift allows us to revisit the questions that motivated Searle and Guala — the stability of institutions and the coordination of social action — without relying on problematic assumptions about mental content.

While there is disagreement – as discussed above – about whether representations should even be taken as tangible, causally significant phenomena, I will argue that for the sake of a foundation for empirical research (not just theoretical accuracy) in social ontology, representations can be conceptualized through the predictive coding/Free Energy Principle framework. This framework treats cognition as simultaneously unfolding between agent and environment and unifies action and perception into rolling sensorimotor cycles. As Friston et al. (2011) put it:

> In this picture of the brain, neurons represent both cause and consequence: They encode conditional expectations about hidden states in the world causing sensory data, while at the same time causing those states vicariously through action. [...] In short, active inference induces a circular causality that destroys conventional distinctions between sensory (consequence) and motor (cause) representations. (p.138)

What this framework provides for social ontology is a way to explain how institutions can be at once created and inspire creation in individuals; at once changing and also seemingly stable; at once constraining and enabling behavior. It has the potential to solve the 4 issues within the ontological frameworks highlighted in section II.

IV. The Free Energy Principle and Representation

IV.1. Introduction

In this section, I will explain the Free Energy Principle (FEP)/predictive coding framework, provide some context for why it is gaining traction in many areas of cognitive science, and provide some examples of the growing body of evidence from neuroscience that points to the FEP and Bayesian prediction as better alternatives to the internalist, feed-forward, contenful conception of representation. I will also provide an in-depth explanation for why the FEP is a good framework for reconceptualizing representations in social ontology Conceptualizing representation through the FEP can help address the four issues in the ontologies of Guala and Searle. I will also provide examples from theorists who have used the framework to formulate alternative explanations for how social institutions function and evolve over time. In order to make clearer the connection between mental representation and social ontology, let us start by first examining how social cognition bridges the gap between individuals and the social, institutional world, along with how the FEP has impacted research in social cognition.

IV.2. Social Cognition and the Free Energy Principle

Some research in social cognition is aimed at addressing how our minds and bodies accomplish the feat of producing complex, stable social phenomena: how exactly do different individuals cooperate and solve complex problems, establish conventions, create hierarchies, make promises, etc. on a cognitive level? As mentioned earlier, one of the prevailing theoretical frameworks for social coordination is shared intentionality (Tomasello & Carpenter 2007; Bratman 1993; Gilbert 1987, 2009; Searle 2010; Tuomela 1995, 2007). Shared intentionality refers to two or more actors cooperating or collaborating, with both their mental processes directed toward the same thing, with both actors aware that their intentions are shared (Tomasello et al. 2005). This framework is important for discussing institutions – it posits the existence of a set of cognitive capacities, allowing for social interaction on a level that does not exist in other animals (Tomasello 1999). It provides a solid foundation for uniquely human social interactions (such as institutions) to exist, albeit with the help of cultural and ontogenetic processes enabled by it (Tomasello 1999). This general framework has since been extensively researched and extended, providing more and more specifics for how it enables the cognition that brings about institutional phenomena.

For instance, Azaad, Knoblich & Sebanz (2021) have written a comprehensive summary on social cognitive abilities, with a specific focus on how these abilities might be framed according to perception and action. Firstly, social coordination requires agents involved to predict each other's actions in order to coordinate successfully (Sebanz & Knoblich 2009). This process of prediction has been researched extensively, with the direct-matching account (Rizzolatti et al. 2001) becoming a prevailing viewpoint, which posits that our perceptual representations of others' actions are matched to our own motor representations, facilitated by mirror-neurons. Azaad et. al point to a body of research highlighting that there is certainly a connection between perceiving actions and motor regions in the brain. The main issue with this account, however, is that it has no way of disambiguating the same action in different contexts. For instance, reaching with an extended finger could be construed as either pointing or pressing, depending on the context. Motor brain regions also don't seem to respond equally to actions in different contexts: Pomiechowska and Csibra (2017) have found that mirror neurons respond to instrumental (reaching), but not communicative actions (pointing). So the role of mirror neurons and perception-action matching remain uncertain in social coordination.

This is where the predictive processing framework proves to be explanatorily useful. One way to reframe this connection between action and perception is through the predictive processing account:

> [Predictive processing posits] that we identify concurrent actions rapidly by forming and updating hypotheses about them as they unfold (Wilson & Knoblich 2005), often with the help of contextual cues such as an actor's environment, dispositions, or explicit intentions (Bach et al. 2014; Bach & Schenke 2017; Csibra 2008; Donnarumma et al. 2017). According to predictive accounts, brain activation during action observation reflects a simulation of the actions that we predict others will perform (Csibra 2008). These simulations might play a role in 'monitoring' whether a seen action is consistent with our predictions. (p. 5-6)

These cues and contexts involve actors' temporal alignment of actions – interpersonal synchrony (Miles et al. 2009) –, interactive cues (Gray et al. 2017), communicative cues (Manera et al. 2010), and increased focus on attention coordination (Vestner et al. 2019), all of which influence actors' perceptions of their environment and their social interactions. Different cues in the environment color our perceptions which also elicit matching motor representations. The line between perception and action becomes blurred and the process of social interaction thus has to be understood in terms of the reflexive effects of perception and action on each other.

This is why Azaad et al. (2021) describe social interactions as having affordances. Since our perception is directly linked to action, the environment becomes colored according to our possibilities for interaction within it. Our interactive relationship with the environment is structured through affordances – perceived possibilities for interaction. Affordances modulate our attention, and attention in social contexts is modulated through a variety of mechanisms that facilitate coordination between agents (Veissière et al. 2019). Such mechanisms include perspective-taking, gaze-following, attending to goal-relevant information in the environment, or temporal synchronizing with other agents. With affordances and attention closely linked, the actions that follow facilitate coordination and are also based in predictability. Importantly, Azaad et al. highlight that actors in a coordination predict and monitor actions in the context of a joint goal. When the parameters of a joint task are vague, agents instead rely on coupled prediction (Noy et al. 2011). In such improvisatory coordination, coupled prediction – trying to predict each other's actions directly – involves each agent being more sensitive to the other's actions, thus forming a coordinative act with no clear leader or follower. This reflexive way of prediction leads to better results in coordination, than simple leader-follower strategies.

All in all, Azaad et al. outline coordination as a process of establishing joint intention and monitoring actions and action possibilities through attention, in context of environmental cues and individual and dyad (or group) level affordances. The modulation of action, perception and attention all together reinforce each other in service of the facilitation of coordination. Agents modulate perception to hone in on affordances, which immediately recruit appropriate actions, which are monitored and adjusted according to incoming perceptive signals (I will return to this point later in more detail). The process is fluid, difficult to dissect, and exists in the context of cooperation. It is thus vital to construe the cooperative situation as a joint act, as the cognitive cues and estimations of individuals within the coordinative act depend partly on the predicted goals. The picture that the authors paint of social cooperation is one where predicting and being predictable are vital. Without a joint goal on which to base predictions on, cooperation could not work. So how exactly does this fluid process of attention, perception, and action in service of prediction work? And what does this mean for representations? In order to imagine representations in the context of prediction-action, we have to establish why its advantageous not just for humans, but for all living things to predict their environment and respond with prediction-appropriate actions.

IV.3. Interlude on Dynamical Systems

Here, we can turn to predictive coding/FEP. Predictive coding as a general theory of cognition gets its basis in the free energy principle (FEP). It is distinct from predictive processing, which is a Bayesian account of brain function, where the main objective of the brain is prediction error minimization, or PEM (Hohwy 2013; Clark 2016). It characterizes the brain as unique in this endeavor when compared to the rest of the organism. In contrast, predictive coding, which has at its center the FEP framework, extends the prediction-and-action interaction to the entire organism. FEP characterizes the ways in which an organism (be it a single cell, a plant, or an animal endowed with a nervous system) exists, and is deeply rooted in dynamical systems theory. More specifically, FEP posits that biological systems are dynamical systems which put an upper bound on the free energy within their systems (Friston 2010).

So what is meant by biological systems being dynamical systems? A complete answer to this question is beyond the scope of the thesis (and certainly beyond the scope of my knowledge of mathematics), but some ideas from dynamical systems theory and self-organization will serve as an extremely useful guide to understanding why predictive coding is so integral to the way biological systems function. As a starting point, it is important to understand that dynamical systems theory arose from the subject of dynamics. As Strogatz (1994) puts it: "This is the subject that deals with change, with systems that evolve in time. Whether the system in question settles down to equilibrium, keeps repeating in cycles, or does something more complicated, it is dynamics that we use to analyze behavior" (p. 2). One contribution to the study of dynamics was made by the Dutch scientist Christiaan Huygens. Huygens discovered the strange properties of pendulum clocks: namely, that two clocks hanging in his room whose pendulums were not synchronized, would eventually come to swing in unison, suspended from the same base (Buyse 2017). The discovery was made by accident, but Huygens was so fascinated with the "odd kind of sympathy" between the two clocks that he began investigating the phenomenon, now known as "spontaneous synchronization". While at first Huygens thought the air was responsible for the spontaneous synchronization of the two clocks, he later discovered through experiments with pendulum clocks that by hanging them on the same base, the pendulums created tiny, imperceptible perturbations in the base, thus slowly influencing each others' swing timings ever so slightly. Such synchronization will be important in the discussions to come, so I will return to it later.

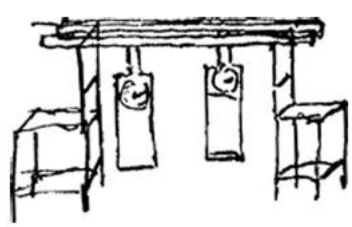


Figure 2 (Buyse 2017, p. 121)

A pivotal contribution to the field came again in the 1600s, from Huygens' contemporary, Sir Isaac Newton, who introduced differential equations, formulated his laws of motion, and developed the theory of universal gravitation. In particular, Newton successfully tackled the two-body problem (a dynamical system involving two bodies and their movement), which involved calculating Earth's motion around the Sun based on the law of gravitational attraction. These breakthroughs allowed him to elucidate Johannes Kepler's laws of planetary motion. Newton's solution was a breakthrough in both physics and mathematics, breeding centuries of effort from mathematicians to apply Newton's solutions to the three-body problem, which surprisingly turned out to be exponentially more difficult.

Later developments in dynamics yielded important insights as to why such problems as the three body problem were so difficult. In the late 19th century, Poincaré's geometric approaches yielded the first ideas of "chaos". In chaos, "a deterministic system exhibits aperiodic behavior that depends sensitively on the initial conditions, thereby rendering long-term prediction impossible" (Strogatz 1994, p. 3). Despite their determinism, such systems can appear "random" and are ultimately fully dependent on their initial conditions. Chaos would later garner increased importance as an element in the field of dynamics, with Edward Norton Lorenz discovering "chaotic motion on a strange attractor", by studying the weather – a notoriously unpredictable, yet seemingly patterned system (p. 3). Lorenz's "solutions to his equations never settled down to equilibrium or to a periodic state – instead, they continued to oscillate in an irregular, aperiodic fashion" (p. 3). With the newfound power of computer simulation, he found that starting his simulations from slightly different initial conditions, resulted in the system behaving in radically different, unpredictable ways very quickly. It turned out that taking measurements from such systems (be it the weather, animal populations, fluid dynamics, or social systems), would serve as shaky ground to build future predictions on. The smallest error in measurement would lead to huge discrepancies between predictions, and the actual behavior of the system.

With the concept of chaos firmly established in the field, starting in the 1970s, Robert May found it in population biology, and used nonlinear equations to study populations and genetics. In "Simple models with very complicated dynamics" (May 1976), for instance, he argues that the dynamics of chaos, along with an array of dynamical behavior – including stable points and bifurcating hierarchy of stable cycles – are pervasive in population biology and genetics. May also argued for a general introduction of nonlinear thinking and nonlinear difference equations to general education.

Systems, such as the universe, our bodies, or a car, can be described as linear or nonlinear.[6] Linear systems display an important characteristic that sets them apart from nonlinear systems: superposition (or additivity and homogeneity, equivalently) (Thyagarajan & Kalpana 2021). Superposition is the principle that in all linear systems, the net response caused by two or more stimuli is equal to the sum of the effects that would be individually caused by each stimulus (Illingworth 1991). This is additivity: essentially, if stimulus X has effect A, and stimulus Y has effect B, then (X + Y) produce (A + B). Homogeneity, on the other hand, is that for any input/stimulus, the scaling of the input always results in scaling the zero-state response by the same factor (Roberts 2018). What this means is that linear systems display behaviors that, derived from a complex input, can still be described as a *sum* of responses to simpler inputs. Linear equations can be used to describe linear systems, and their solutions can be combined to get another solution (Barton 1994). Nonlinear systems, by contrast, are a bit more complicated.

The aforementioned mathematical properties make modeling linear systems much simpler than many nonlinear systems. The superposition principle does not apply to nonlinear systems and thus, such systems display behaviors that are not proportional to initial inputs, which is why their behaviors cannot be predicted without knowing the original inputs. This peculiar property of nonlinear systems is what makes them so difficult to accurately model. Nonlinear equations can be used, but their solutions are often difficult to solve for and give a pattern of solutions. An example of this is the equation $x^1 = x + rx (1 — x)$, which is type of nonlinear equation called iteration. It takes each consecutive solution and plugs it into the equation again and again. The splits seen in the figure occur suddenly, at specific values of *r*. The iterations of

---

6 Technically, systems are linear or nonlinear depending on our frame of reference. Some systems are described as linear since, for the purposes and time frame of our interactions with them, they display linearity. For instance, a computer may decay over the course of a few years, but during those years, it behaves predictably enough so that I can perform certain actions on it and reasonably expect it to give me outputs that fit my expectations. Without maintenance, however, it would eventually stop functioning predictably.

the solutions turn into a pattern, forming bifurcations that eventually turn into chaos. The equation demonstrates the close relationship between chaos and nonlinearity, as the can be seen in Figure 3 (Barton 1994).

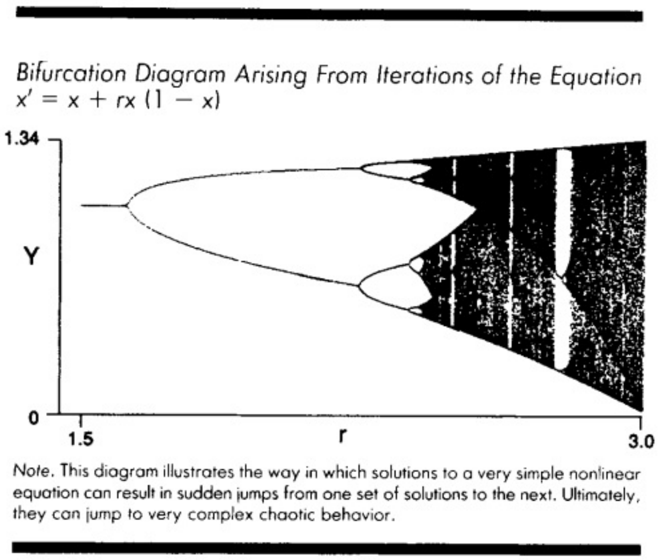


Figure 3

Nonlinear systems tend to settle down over time. They do this in a very particular way, forming into patterns that exhibit periodic behavior, called attractors. These attractors tend to converge onto a discrete point, a simple oscillating cycle, a quasiperiodic cycle, or a chaotic cycle. These can be seen in Figure 4 (Barton 1994). Chaotic attractors require a special sort of explanation:

> The pattern in such an attractor is bounded, but after a certain number of repetitions within the system, it becomes very irregular. This irregularity results in unpredictability, despite the fact that it derives from a completely deterministic system. This unpredictability is associated with a property of chaotic systems known as sensitive dependence on initial conditions. This means that if two sets of initial conditions differ by any arbitrarily small amount at the outset, their specific solutions will diverge dramatically from one another over the long range. In the case of the pattern in [Figure 4], for instance, if x = 0.5000 had been replaced with x = 0.5001, the solutions in the chaotic portion of the attractor would have been completely different, whereas the general chaotic pattern would be the same. Given that no measurement system is without some error, it becomes clear that if a system is chaotic, general patterns of future behavior may be predictable but specific behaviors over the long range will not. (p. 6)

Types of Attractors Typically Found in Nonlinear Systems: (A) Point Attractor, (B) Cyclical or Oscillating Attractor, (C) Quasiperiodic Attractor, (D) Chaotic Attractor

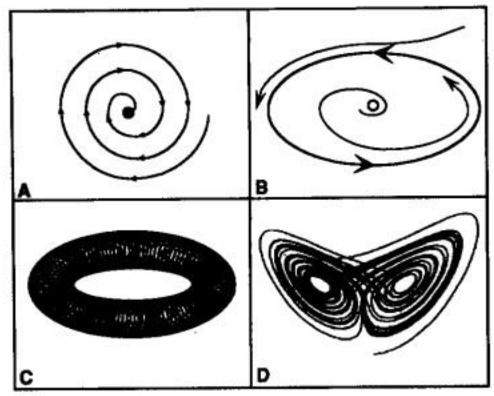


Figure 4

Systems such as three bodies rotating around each other, weather patterns, brains, and societies are fundamentally “chaotic”, while also displaying recurring cycles, patterns, and oscillations. Robert May (1976) himself felt that the hegemony of linear thinking in elementary mathematics courses at university gave students an unrealistic representation of the world, as “nonlinear systems are surely the rule, not the exception, outside of the physical sciences” (p. 93).

IV.4. Dynamical Living Systems and the FEP

Now we can return to our original issue: how the study of dynamics can be used to describe living systems. As we’ve seen from the research of Poincaré, May, and Lorenz, from the often random and chaotic nature of dynamical systems, there are recurring patterns and oscillations that ebb and flow in and out of existence. As the universe in its entirety can be looked at as one massive dynamical system, it is the milieu in which living systems must come about (Strogatz 1994; Tsacher & Dauwalder 2003). Living systems being systems themselves, ones which somehow emerge and endure for some time, have to organize themselves into patterns that can withstand the general tendency of the universe toward entropy or chaos (Friston

2010; Bruinberg et al. 2018). This happens through the dynamical process of self-organization.

Willaim Ashby (1947), who originally formulated the concept of self-organization, describes self-organization as closely linked to nonlinear dynamical systems. Upon reaching an attractor, a self-organizing system becomes bound to continue its development within this state. This restriction signifies a type of interdependence or synchronization among its individual components or subsystems. Each subsystem has adjusted itself to the environment created by all the other subsystems. Bonabeau, Dorigo, and Theraulaz (1999), who examine swarming insects to better understand self-assembly, offer another perspective: self-organization “is a set of dynamical mechanisms whereby structures appear at the global level of a system from interactions among its lower-level components” (p. 9). They claim that self-organization, by definition, carries with it the seeds of intelligence. Self-organizing systems are decentralized problem-solving systems, that solve the problem of their own continued existence against the pull of entropy (or total chaos, in DST terms). Self-organization is reliant on several processes, one of which is the amplification of fluctuations – self-organization happens not just against chaos, but through it. A dynamical system’s random fluctuations allow for new solutions and patterns to emerge, which are either antithetical or supportive of the system’s dynamics – as Bonabeau and colleagues put it, positive and negative feedback. This happens over multiple interactions through time and space between the components of the system, and require energy to continually stave off entropic tendencies.

What living systems accomplish then, is to stay within a certain set of states, which constitute their continued existence – they are nonlinear dynamical systems that evolve toward this set of states that can be described as an attractor (Friston et al. 2014). These attractor states are the states that a living system converges toward and are the states in which it can be said to be “alive” (Friston 2010; 2011). The mechanisms behind this dynamical process can be formulated according to the free-energy principle (FEP), describing how these self-organizing systems differentiate, produce, and maintain themselves (Friston 2010; 2018; Wiese & Metzinger 2017; Badcock, Friston & Ramstead 2019). FEP is a potential unifier in theoretical neuroscience and theoretical biology, along with social cognition (Bruineberg et al. 2018). Most simply put:

> FEP […] leverages the simple truth that agents that exist do so because they are able to persist, appearing to resist – for periods of time – the second law of thermodynamics that states that entropy (disorder) increases over time. Biological agents are able to resist because the second law applies only to isolated (or closed) systems. By exchanging matter and energy with the environment, such systems are able to preserve their own integrity and order. They do so, of course, only by increasing disorder elsewhere (thus 'obeying' the second law). We thus enter the realm of living or adaptive systems – systems that actively seek out, and work to bring about, the conditions that are necessary for their own survival. (Clark 2017, p. 3)

FEP is thus a description of how such biological systems operate. These systems can be as small as single cells or as complex as clovers, Komodo dragons, humans, or forests.[7] Although the systems vary in size and organization, they all aim to put a bound on free energy.

It is important to emphasize that these systems are not isolated systems; they exist coupled with their environment. As Friston (2010) puts it, any system that is in equilibrium with its environment must minimize free energy. In this sense, the delineation of the self (of the agent) happens not just against the environment, but with it. By minimizing free energy, agents not only create and maintain themselves, but also modulate and are modulated by the states of the environment that they exist in. This modulation creates a dynamical equilibrium (with both the agent and its environment being dynamical systems), much like the synchronization of Huygens' pendulum clocks. For Huygens, the inner states of one clock adjusted to the outer states of the other clock through their shared base, much like how the inner states of an agent adjust to the outer states of their environment, per the Markov blanket they share (a cell wall, mucus membrane, skin, etc. I will return to this idea).

Because the systems vary in size and organization, their equilibrium with the environment is also unique. This makes autopoiesis a unique endeavor for each living system, based partly on their internal dynamics and the dynamics of the environment. FEP

> formalizes the self-maintenance aspect of autopoiesis in terms of Shannon information: of all possible states the organism could find itself in, the organism

---

[7] These organisms are complex systems characterized by non-linear interactions among multiple components that are hierarchically arranged and nested (Hilgetag & Goulas 2020; Kauffman, 1993, 1995; Kirchhoff et al., 2018). Understanding their self-organization requires not only examining how individual components interact but also how processes at microscopic and macroscopic levels influence each other, which involves both top-down effects and bottom-up influences.

must find the right subset of states that allow its organization to persist in its energetic exchanges with the environment. The claim is that there is a probability distribution over all possible states the organism can find itself in. At any point in time, this distribution is sharply peaked around certain values specifying conditions of the organism that are necessary for its viability and survival or – more exactly – that define the characteristic phenotypic state of the organism. (Bruinberg et al. 2018, p. 2421)

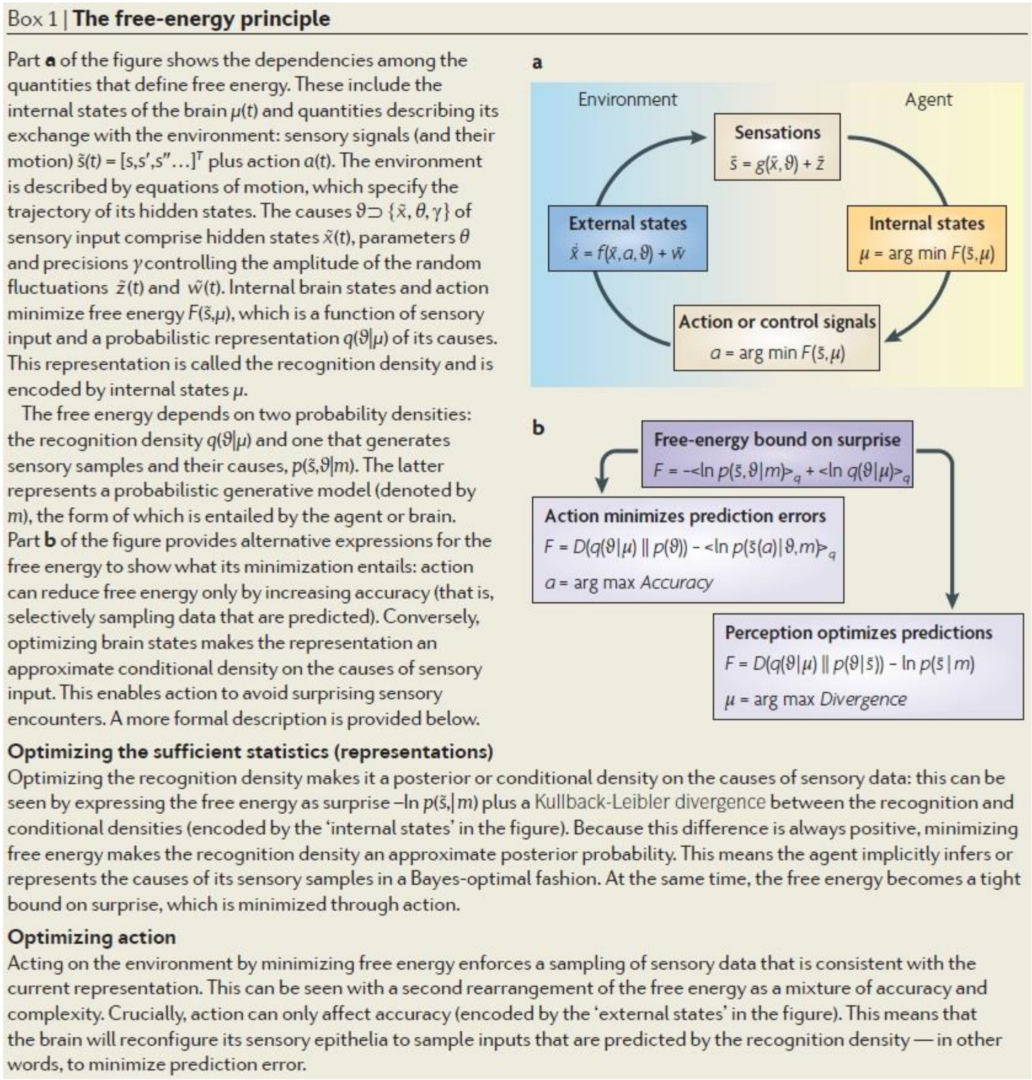
It is important to define, at this point, how exactly the living system separates itself from its environment. Both the living system and its environment are dynamical systems and their separation can be described as statistical (Friston 2013; Bruinberg et al. 2018; Clark 2016). As hinted to earlier, this statistical separation comes in the form of a Markov blanket. The Markov blanket separates the inner states of a system, which are statistically independent of the outer states of the environment, while the Markov blanket itself consists of the states that separate the two (Friston 2010; Clark 2016; Kirchhoff et al. 2018). What is meant by states is: "any variable that locates the system at a particular point in state space; for example, the position and momentum of all the particles constituting a thermo-dynamic system – right through to every detail of neuronal activity that might describe the state of the brain" (p. 1). In the case of Huygens' clocks, each clock is a dynamical system, separated by a base (a wall, a wooden beam, or whatever they are suspended on). From the perspective of one clock, the other clock's states are unknown and inaccessible to it, as the two are separated by a Markov blanket (the base on which both clocks are suspended). Despite being statistically independent, the states of both clocks become more aligned (their pendulums synchronize) through their interactions with the states of the Markov blanket, their shared barrier.[8]

[8] If Markov blankets are the crucial barrier around which living systems create and sustain themselves, how come it can be present in inanimate matter as well? There is a crucial difference between the Markov blankets in Huygens' clocks and the Markov blankets of a brain or a pine tree: The behavior of Huygens' pendulums demonstrates a time-dependent Markov chain, which is a simplified form of a Markov blanket. In a Markov chain, states follow a linear progression, with each state influencing the next without any reciprocal or looping effects. This creates unidirectional dependencies, meaning that events at a given time are conditionally independent of both past and future events, given the current state. Systems described by Markov chains are non-autonomous since they cannot influence or adjust their interactions with the environment; they are constrained by their immediate circumstances and past conditions. In contrast, biological systems differ significantly. They maintain homeostasis across multiple time scales, allowing them to respond dynamically to environmental changes. For this, they require a generative model that predicts future states and enables them to choose actions with the highest

To reiterate, both the agent and the environment it finds itself in, are dynamical systems. The agent, or biological system, avoids entropy by separating itself from its environment (which naturally tends toward entropy); this partition can be described using a Markov blanket. Thus, the free energy principle can be thought of as the mechanism that explains the separation and attunement/misattunement between the dynamics of these two systems, through their interactions with the Markov blanket. As stated previously, according to Friston's (2010, 2011) conception, free energy is the function of several quantities: the organism's inner dynamics, its "sensory states" (which are the states of the Markov blanket) and its "active states" (also a part of the Markov blanket). Free energy is thus an estimation of the constant misattunement between the distal causes of the sensory states (the environment) and the inner states of the biological system, and its minimization can work both in the direction of sensory states (perception) or toward active states (action). Through this attunement, the inner states of the biological system are maintained within a viable set of physiological bounds – also known as allostasis (Sterling & Eyer 1988; Sterling 2012). This interaction can be seen in more detail in Figure 5 (Friston 2010).

---

expected evidence or lowest free energy. Biological systems, therefore, exhibit much more adaptive and flexible behavior than a simple Markov chain. (Kirchhoff et al. 2018)



Box 1 | **The free-energy principle**

Part **a** of the figure shows the dependencies among the quantities that define free energy. These include the internal states of the brain $\mu(t)$ and quantities describing its exchange with the environment: sensory signals (and their motion) $\tilde{s}(t) = [s, s', s'' \ldots]^T$ plus action $a(t)$. The environment is described by equations of motion, which specify the trajectory of its hidden states. The causes $\vartheta \supset \{\tilde{x}, \theta, \gamma\}$ of sensory input comprise hidden states $\tilde{x}(t)$, parameters θ and precisions γ controlling the amplitude of the random fluctuations $\tilde{z}(t)$ and $\tilde{w}(t)$. Internal brain states and action minimize free energy $F(\tilde{s}, \mu)$, which is a function of sensory input and a probabilistic representation $q(\vartheta|\mu)$ of its causes. This representation is called the recognition density and is encoded by internal states μ.

The free energy depends on two probability densities: the recognition density $q(\vartheta|\mu)$ and one that generates sensory samples and their causes, $p(\tilde{s}, \vartheta|m)$. The latter represents a probabilistic generative model (denoted by m), the form of which is entailed by the agent or brain. Part **b** of the figure provides alternative expressions for the free energy to show what its minimization entails: action can reduce free energy only by increasing accuracy (that is, selectively sampling data that are predicted). Conversely, optimizing brain states makes the representation an approximate conditional density on the causes of sensory input. This enables action to avoid surprising sensory encounters. A more formal description is provided below.

Optimizing the sufficient statistics (representations)

Optimizing the recognition density makes it a posterior or conditional density on the causes of sensory data: this can be seen by expressing the free energy as surprise $-\ln p(\tilde{s}|m)$ plus a Kullback-Leibler divergence between the recognition and conditional densities (encoded by the 'internal states' in the figure). Because this difference is always positive, minimizing free energy makes the recognition density an approximate posterior probability. This means the agent implicitly infers or represents the causes of its sensory samples in a Bayes-optimal fashion. At the same time, the free energy becomes a tight bound on surprise, which is minimized through action.

Optimizing action

Acting on the environment by minimizing free energy enforces a sampling of sensory data that is consistent with the current representation. This can be seen with a second rearrangement of the free energy as a mixture of accuracy and complexity. Crucially, action can only affect accuracy (encoded by the 'external states' in the figure). This means that the brain will reconfigure its sensory epithelia to sample inputs that are predicted by the recognition density — in other words, to minimize prediction error.

Figure 5

As biological systems minimize free energy, they update their inner states to be in synchrony with the outer states of the environment or act on the outer states (through the Markov blanket) to bring them into synchrony with their inner states. As previously mentioned, since there is a Markov blanket between the inner states of the agent and the outer states of the environment, there is a statistical separation between them. Thus, in order to extract information from its environment, each biological system must contend with the question: what are the causes in the environment, whose effects it detects in the states of the Markov blanket? In other words, what are the causes of my sensory data? The agent has no direct access to ever-changing outer states, only to the sensory states of the Markov blanket, so it must infer the causal factors that produce changes in the Markov blanket – this is known as an inverse problem. The minimization of free energy, therefore, involves somehow solving this inverse problem.

In free energy minimization, free energy becomes a value that is as an upper limit which biological systems can access in order to limit their own disorder (Friston 2010; 2011; 2018). It serves to constrain the possible states the organism can find itself in, while also modulating its internal model. By modeling the possible causes of changes in the Markov blanket, a biological system can statistically calculate how best to anticipate the unknown dynamics of the environment and how best to maintain attunement with them. The key importance of free energy is that it is variational, meaning it can be formulated in three different ways (for mathematical details, see Figure 6). This variational aspect of FEP is what allows for organisms to at once estimate their own borders, maintain their inner states to maintain their borders, and regulate exchanges with the environment, thus resisting entropy, and continually create and integrate higher and lower levels of nested Markov blankets to function in synchrony. So what are the three formulations of free energy and what do they mean?

IV.5. Variational Free Energy

First, energy minus entropy: this initial expression establishes a connection between the notion of free energy as employed in information theory and the concepts utilized in statistical thermodynamics – more specifically, Shannon entropy and Helmholtzian entropy.[9] It also

[9] Shannon entropy is concerned with the uncertainty in a set of possible events or outcomes (Shannon 1948). In communication theory, it measures the unpredictability or information content of a message or a system. Higher Shannon entropy indicates greater uncertainty about which event will occur next. Helmholtzian entropy, in thermodynamics, refers to the dispersal of energy and is a component of Helmholtz free energy, which represents the energy available to do work in a system at constant temperature and volume (Helmholtz 1882). Helmholtzian entropy essentially describes how energy spreads out in a system, with systems naturally moving towards states of higher entropy or disorder. The connection between the two lies in their roles in describing systems' tendencies toward equilibrium or predictability. In statistical mechanics, where thermodynamics meets information theory, Shannon entropy can be interpreted as the amount of uncertainty in a system's microstates, akin to the way Helmholtzian entropy describes the system's distribution of energy across states. In both cases, entropy reflects the system's tendency toward disorder or equilibrium (Jaynes 1957).

demonstrates that an agent can assess free energy because the energy component represents the unexpectedness regarding the simultaneous occurrence of sensations and their perceived causes, while entropy simply

Supplementary information S2 (box): Variational free energy

Here, we derive the free-energy and show how its various formulations relate to each other. We start with the quantity we want to bound; namely, the surprise or log-evidence associated with sensory states $\vec{s}(t)$ that have been caused by some unknown quantities $\vartheta \ldots \{\vec{x}, \theta\}$, which include the hidden states and parameters in box (S1)

$$-\ln p(\vec{s}(t)) = -\ln \int p(\vec{s}(t), \vartheta) d\vartheta \qquad \text{S2.1}$$

To create a free-energy bound on surprise $\mathsf{F}\ (\vec{s}(t), q(\vartheta))$, we simply add a non-negative cross-entropy between an arbitrary (recognition) density on the causes $q(\vartheta)$ and their posterior density $p(\vartheta \mid \vec{s})$ (dropping the dependency on m for clarity).

$$\begin{aligned} \mathsf{F} &= \int q(\vartheta) \ln \frac{q(\vartheta)}{p(\vartheta \mid \vec{s})} d\vartheta - \ln p(\vec{s}) \\ &= D(q(\vartheta) \,\|\, p(\vartheta \mid \vec{s})) - \ln p(\vec{s}) \end{aligned} \qquad \text{S2.2}$$

The cross-entropy term is non-negative by Gibb's inequality. In short, free-energy is *cross-entropy plus surprise*. Because surprise depends only on sensory states, we can bring it inside the integral and use $p(\vartheta, \vec{s}) = p(\vartheta \mid \vec{s}) p(\vec{s})$ to show free-energy is *expected energy minus entropy*

$$\begin{aligned} \mathsf{F} &= \int q(\vartheta) \ln \frac{q(\vartheta)}{p(\vartheta \mid \vec{s}) p(\vec{s})} d\vartheta \\ &= \int q(\vartheta) \ln q(\vartheta) d\vartheta - \int q(\vartheta) \ln p(\vartheta, \vec{s}) d\vartheta \\ &= -\langle \ln p(\vartheta, \vec{s}) \rangle_q - \langle -\ln q(\vartheta) \rangle_q \end{aligned} \qquad \text{S2.3}$$

where $-\ln p(\vartheta, \vec{s})$ is Gibb's energy. A final rearrangement, using $p(\vartheta, \vec{s}) = p(\vec{s} \mid \vartheta) p(\vartheta)$, shows free-energy is also *complexity minus accuracy*, where complexity is the cross-entropy between the recognition $q(\vartheta)$ and prior density $p(\vartheta)$

$$\begin{aligned} \mathsf{F} &= \int q(\vartheta) \ln \frac{q(\vartheta)}{p(\vec{s} \mid \vartheta) p(\vartheta)} d\vartheta \\ &= \int q(\vartheta) \ln \frac{q(\vartheta)}{p(\vartheta)} d\vartheta - \int q(\vartheta) \ln p(\vec{s} \mid \vartheta) d\vartheta \\ &= D(q(\vartheta) \,\|\, p(\vartheta)) - \langle \ln p(\vec{s} \mid \vartheta) \rangle_q \end{aligned} \qquad \text{S2.4}$$

Figure 6 (Friston 2010)

reflects the agent's recognition density. In addition, it underscores that free energy relies on a generative model of the world, characterized by the probability of both a sensation and its causes occurring together. Consequently, an agent must possess an implicit generative model depicting how causes collaborate to generate sensory data, shaping both the agent's essence and the quality of the free-energy constraint on surprise.

Second, free energy can be expressed as the sum of surprise and a divergence term. The perceptual divergence, in this context, denotes the distinction between the recognition density and the conditional density (or posterior density) of the causes of a sensation, given the sensory signals. This conditional density signifies the most accurate estimation of the true causes. The discrepancy between these two densities is consistently non-negative, making free energy an upper limit on surprise. Consequently, minimizing free energy by adjusting the recognition density (while keeping sensory data constant) diminishes the perceptual divergence, thereby aligning the recognition density with the conditional density and transforming free energy into surprise.

Lastly, free energy can also be expressed as complexity minus accuracy, utilizing model comparison terminology. Complexity, often referred to as Bayesian surprise, delineates the distinction between the recognition density and the prior density on causes. It represents the contrast between prior beliefs about the state of the world before assimilating sensory data and posterior beliefs encoded by the recognition density. Accuracy, on the other hand, reflects the surprise associated with sensations that are anticipated under the recognition density. This formulation illustrates that minimizing free energy by altering sensory data (while keeping the recognition density unchanged) necessitates an enhancement in the accuracy of an agent's predictions. In essence, the agent will selectively sample sensory inputs that align with its expectations, a concept known as active inference (I will return to this concept in more detail).

This third formulation can be used as a description of the brain, in which the brain is a model of its “outside world” (its prior density), which it contrasts with the incoming sensory data encoded by the recognition density (its sensory signals), which results in a posterior density. In this case, the “outside world” is both the environment of the agent, and the bodily system that surrounds the brain. In this formulation,

> the brain is an inference machine that actively predicts and explains its sensations. Central to this hypothesis is a probabilistic model that can generate predictions, against which sensory samples are tested to update beliefs about their causes. This generative model is decomposed into a likelihood (the probability of sensory data, given their causes) and a prior (the *a priori* probability of those causes).

> Perception then becomes the process of inverting the likelihood model (mapping from causes to sensations) to access the posterior probability of the causes, given sensory data (mapping from sensations to causes). This inversion is the same as minimizing the difference between the recognition and posterior densities to suppress free energy. Indeed, the free-energy formulation was developed to finesse the difficult problem of exact inference by converting it into an easier optimization problem. (Friston 2010, p. 129)

Perception becomes a constructive process based on internal, generative models and incoming sensory information, resulting in the modeling of outer states using hierarchical Bayesian inference. This Bayesian modeling is a natural extension of the free energy principle, and is usually referred to as predictive coding, as the purpose of this model is to compare predictions to incoming sensory data, minimizing the difference between the two (Clark 2015; Friston 2010; 2011; Brown & Brüne 2012; Kilner, Friston & Frith 2007).

IV.6. Cognition through the lens of FEP

Thus, brains can be said to

> minimize (on average) their free energy, where free energy is an upper bound on (negative) marginal likelihood or evidence (i.e. the probability of finding the system in a particular state, given the system in question). This implies that the kind of self-organization of Markov blankets we consider results in processes that work entirely to optimize evidence, namely self-evidencing dynamics underlying the autonomous organization of life, as we know it. In Bayesian statistics, the evidence is known as 'model' evidence" (Kirchhoff et al. 2018, p. 2).

The variational Bayes approach to explaining cognition is to frame brains as inference machines that embody a probabilistic, generative model. This model (Markov blanket inner states) compares its predictions against incoming sensory information to better understand (predict) their underlying causes. In this sense, brains and biological systems can be thought of as "self-evidencing", predicting their own states, embodying a model of the environment's regularities.

As mentioned previously, the Bayesian formulation of free energy involves complexity and accuracy. Thus, brains embody the "infomax principle" (Friston 2010). In it,

> mutual information is optimized when conditional expectations maximize accuracy (or minimize prediction error), and efficiency is assured by minimizing complexity. This ensures that no excessive parameters are applied in the generative model and leads to a parsimonious representation of sensory data that conforms to prior constraints on their causes. (p. 131)

Brains are in the business of survival and so must oscillate between low-complexity modeling – which saves processing power and allows fast decisions at the price of accuracy – and high-complexity, accurate modeling, which requires more computing and thus slows decision-making speed. This way, brains ensure that they're expending the least amount of energy on calculations while maximizing the accuracy of their predictions. It is intriguing that sophisticated model optimization methods employ free-energy optimization to remove redundant model parameters. This implies that free energy optimization could offer a compelling explanation for the synaptic pruning and homeostasis observed in the brain during neurodevelopment and sleep (Paus et al. 2008; Gilestro et al. 2009). This is unlike traditional, feed-forward representation-manipulation, where the outside world needs to be represented and stored in the mind and accessed when necessary, requiring virtually unlimited, biologically impossible storage capacities and combinatorial abilities, making the brain slow and inflexible in a fast-changing environment (Hawkins & Blakeslee 2004).

There is also evidence that the brain not only weighs prediction against sensation and initiates action moment to moment, but also itself embodies in its biology an organization guided by the minimization of free energy. Neurons in the brain become synchronized through correlated pre- and postsynaptic activity: neurons that fire together, wire together.[10] The causes of these patterns, however, need to be inferred by the brain. The correlation theory (von der Malsburg 1981), delves into how synaptic efficacy and plasticity are selectively facilitated by rapid synchronous activity induced by various perceptual attributes of the same object. This addresses a limitation of classical plasticity, which struggles to attribute a presynaptic input to a specific cause in the external world. The correlation theory underpins theoretical treatments of synchronized brain activity and its role in associating or binding attributes to specific objects or

[10] In *The Organization of Behavior: A Neuropsychological Theory* (1949), Donald Hebb proposed what is now called "Hebbian learning," a theory explaining how learning and memory could be mediated by the strengthening of synaptic connections between neurons that are simultaneously active. This idea has been summarized by the phrase "neurons that fire together wire together," even though Hebb did not use these exact words. Instead, the phrase encapsulates the core idea that repeated, simultaneous activation of neurons can lead to the strengthening of their synaptic connections.

causes. This optimization process involves adjusting connection strengths between neurons, a mechanism fundamental to learning.

Performing gradient descent on free energy – altering connections to minimize free energy – is formally equivalent to Hebbian plasticity (Friston 2010). This equivalence arises because the parameters of the generative model dictate how anticipated states (synaptic activity) are combined to generate predictions. In essence, when the predictions made by presynaptic activity and the errors in predictions made by postsynaptic activity are strongly correlated, the connection strength increases. The brain strengthens local neural connections that often fire together, thereby also forming predictions that better anticipate long-term causal processes while also selectively enabling synaptic efficacy and its plasticity. What this means is that Hebbian learning strengthens correlations, while selective plasticity allows for learning. This balancing act between strengthening associations and keeping them flexible is underpinned by minimizing free energy: inference on states (perception) and inference on parameters (learning) minimize prediction error, limiting surprising interactions with the environment.

Through synchronization and selective plasticity, brains regulate causal patterns in synaptic efficacy to predict the progression of states within the world. However, brains also have to gauge the precision of the incoming flow of sensory information in the form of prediction error. For example, hearing a single noise in the forest makes it difficult to make inferences about it – in other words, our brain's model for the incoming sensory input does not assign high precision to the noise (given a brain model where there can be lots of things in the forest that make noise, our brain's model does not have enough input to determine the source of the noise). So how does our brain judge the reliability of incoming sensory input? Building on the chaotic nature of dynamical systems mentioned previously, it is actually random fluctuations in sensory states that play a key role in generating sensory data and providing a measure of precision. Their amplitude is the inverse variance, or precision, encoding the reliability of prediction errors. More precisely:

> precision modulates the amplitude of prediction errors, so that prediction errors with high precision have a greater impact on units that encode conditional expectations. This means that precision corresponds to the synaptic gain of prediction error units. The most obvious candidates for controlling gain (and

> implicitly encoding precision) are classical neuromodulators like dopamine and acetylcholine, which provides a nice link to theories of attention and uncertainty. Another candidate is fast synchronized presynaptic input that lowers effective postsynaptic membrane time constants and increases synchronous gain. This fits comfortably with the correlation theory and speaks to recent ideas about the role of synchronous activity in mediating attentional gain. (Friston 2010, p. 132)

Precision is especially important in hierarchical frameworks, since it maintains the balance between the impact of top-down predictions and bottom-up prediction errors. More precisely, the enhancement of expected precision through synaptic gain intertwines attention with synaptic gain and synchronization. This connection plays a pivotal role in theories concerning attentional gain and biased competition, particularly within the realm of neuromodulation.[11] Thus, precision plays a role in ensuring low-quality sensory information does not end up creating inaccurate predictions about the world while also playing a role in attention, which guides sensory sampling and increases precision (and with it synaptic gain), producing prediction errors that have higher impact due to their perceived precision. As per our example, attuning our ears to listen for furthers noises is an attempt to remedy the low precision of our model. If no further noises come, it might be a safe bet that we do not need to revise our model of our surroundings – we are most likely alone and safe.

These different balancing acts that the brain, as free energy minimizing model, performs, allow it to essentially self-regulate its learning processes, detect environmental and bodily regularities, and modulate its attention, all while keeping the energetic costs of these processes at a minimum. It is not enough, however, for the brain to simply "model" and predict the outer

[11] Attentional gain refers to the aspect of the FEP framework that describes the relationship between the precision of predictions and attention (Feldman & Friston 2010). Attention is a mechanism for optimizing perceptual inference by reducing uncertainty. Under the free-energy principle, the brain strives to resolve these errors by allocating attentional resources to particular sensory signals, effectively biasing competition among stimuli. In this way, attention enhances the processing of relevant stimuli while suppressing less relevant or unexpected ones, optimizing the brain's predictive accuracy. In terms of biased competition, Feldman and Friston suggest that attention can be seen as a selective process driven by prediction errors. These prediction errors arise when there is a mismatch between expected and actual sensory input. This extends the concept of biased competition in neuroscience, where different stimuli compete for neural representation, and attention biases this competition toward stimuli that are behaviorally relevant.

states of the environment, as passively anticipating changes in the environment would do little to stop them from overwhelming the allostatic balance that keeps inner states separate from the environment and dynamically stable. What is required is action; in FEP terminology, the process that brings about action is called active inference.

As stated previously, the allostasis that biological systems exist in must be actively maintained through interactions with the Markov blanket. Sensory states are influenced by external states and inform, but cannot be influenced by, internal states, whereas active states are influenced by internal states and cannot be influenced by outer states, but instead influence outer states. Organisms can thus increase the complexity of their model by updating it through error signals coming from sensations, or by making them more accurate by sampling the environment through action, molding the environment to produce sensory information that confirms their predictions – all still within the Bayesian framework. The trick is that instead of minimizing error by updating predictions, predictions drive actions in an attempt to change the environment, thereby decreasing the expected prediction error in perception. In this way, predictions can be thought of as goal states, which the system tends toward. The agent acts on its environment to change the incoming sensory data to better fit its internal predictions (active inference), or it updates its internal predictions to better fit the incoming sensory information (perceptual inference).

IV.7. Perception-action Loops and Active Inference

What may be becoming apparent is that in this Bayesian framework, which balances between complexity and accuracy, perception and action become blurred. The mechanism driving both action and perception – the constant minimization of free energy – produces rolling sensorimotor cycles, where predictions drive sampling behavior that constrain what sensory information the biological system attunes to and interprets (recognition density), which then may update the predictive model (posterior density) based on the likelihood of the data and the prior density (Friston 2013). Alternatively, the system may favor its existing priors and use them to drive further actions to sample further sensory information, and on and on these cycles go. This process points to a circular, looping way that biological systems relate to their environments and their inner dynamics, being molded by and at the same time molding them. By leveraging the two strategies of perceptual inference and active inference, the agent minimizes free energy, thus maintaining its energetic exchanges with the environment, maximizing model evidence for its own existence, taking shape by restricting itself to a set of most likely states, and resisting entropy by maintaining itself in a non-steady state equilibrium by interacting with the environment. Through this process, internal and external states reciprocally cause each other, with this circularity being the fundamental architecture of self-organizing systems (Jörg 2011). A similar circular causality was suggested by Merleau-Ponty, the "intentional arc", where form and function, body and envrionment are tightly bound and are reciprocally dependent on each other.[12]

Active inference takes the inference problem faced by a biological system and flips it on its head. Through action, a living system can influence the data which it will detect through perceptual inference (Friston 2010; 2013; Clark 2016). Instead of adjusting the parameters of the internal model to match the incoming sensory data, the agent can manipulate the environment through its Markov blanket in order to change the states of the environment that better fit the prediction of its internal model. For instance, when I pick up a glass of water, I predict that the glass will be in the air (even though my senses tell me it is currently on the table). In order to minimize the divergence between my prediction (a lifted glass) and my current sensory input (glass on the table) I lift my arm to minimize this divergence, lifting the glass into the air. Predictions for action sequences also come from the internal model.

[12] Maurice Merleau-Ponty's concept of the intentional arc plays a central role in his existential phenomenology, particularly in his major work *Phenomenology of Perception* (1945). The intentional arc refers to the reciprocal relationship between the body, perception, and the world. According to Merleau-Ponty, human experience is not a series of disconnected sensory events but an integrated, meaningful interaction between the body and its surroundings. The intentional arc captures the way our past experiences, body, habits, and perceptions dynamically structure our present engagement with the world. This arc integrates sensory-motor capacities, perception, and higher-order activities such as thinking and willing, guiding us in future actions. The more we engage with the world, the more skilled and attuned we become to it, thus enhancing this intentional arc. This concept is quite similar to the formal description of cognition described by the FEP and enactivism, where body, brain, and environment attune to and mold each other in a perpetual balancing act.

The actions chosen by the agent are always based on its previous interactions with the environment (i.e. its perceptual model) – this is referred to as the Ideomotor Principle (Prinz 1990; Hommel 2015). More specifically, it states that the neural underpinnings of perceptual inference and of action preparation are identical. Computationally, the Ideomotor Principle leverages a formal duality between action and perception. The duality is as follows: If I can perceptually access a state of affairs *p*, it implies that *p* has some perceivable consequences (or constituents) *c*. Action is goal-oriented, meaning that by performing an action, I aim to bring about some state of affairs *p*. Thus, action can be described as the process of producing the perceivable consequences *c* of *p*, and perception can be described as the process of inferring the causes of a hypothetical action that results in *p* and, consequently, *c* (Todorov 2009; Wiese & Metzinger 2017). Thus, a living system chooses action sequences that have the highest potential to minimize free energy according to its model of the world (e.g. I don't choose to levitate the glass with my mind, because my model has no previous evidence for my powers of levitation. It does, however, expect high probability that my arm will be able to lift the glass, making my arm the ideal vehicle to lift the glass). By inferring the causes my sensations, I have also inferred what I should do in order to change those causes and change what I perceive. This is why cognition can be thought of as cycles of sensorimotor loops, where action and perception are tied together to form stable interactions with the environment.

By leveraging action, the agent keeps the interactions between itself and the environment predictable and minimizes the free energy not just of its current state, but the free energy of the dynamic coupling between itself and the environment over the short term and long term (Wiese & Metzinger 2017). This is because the living system is fundamentally in the business of staying within a predictable set of states that allow for its continued existence. Action is a fundamental product not just of inference, but coincidentally of survival, as it allows for the organism to adjust the states of the environment, change the dynamic coupling, and maintain allostasis.

What the FEP framework amounts to is that modeling, surviving, and perception/action are all built around the minimization of free energy and are thus interrelated. It describes a biophysics for biological systems: since linear modeling methods cannot properly predict dynamical systems, a biological system becomes "a physical transcription of causal regularities in its eco-niche that has been sculpted by reciprocal interactions between self-organization and selection over time” (Ramstead et al. 2018, p. 5). The free energy principle “tells us how the quantities that define Markov blankets change as the system moves towards its variational free energy minimum” (p. 4). In other words, the system creates, maintains, and interacts with its own Markov blanket by minimizing free energy. As we build increasing complexity from individual neurons to neuronal clusters to interconnected brain areas – while still adhering to the principle of free energy minimization – it becomes a bit hazy how all this complexity and diversity in size and function can organize around just one principle.

Markov blankets provide the basic building blocks for this complexity. Much like how cells are the basic building blocks for multicellular organisms, Markov blankets form ever-larger ensembles guided by self-evidencing, active inference – In Kirchhoff et al. (2018), this is termed “adaptive active inference”. Organisms consist of numerous Markov blankets, the boundaries of which are malleable and constantly shifting. According to Kirchhoff and colleagues, a collection of these Markov blankets can self-organize into a larger, global, or macroscopic system, which itself possesses a Markov blanket. This concept demonstrates how autonomous systems are formed through multiple layers of self-evidencing and nested Markov blankets. Consequently, this framework suggests that a living system is comprised of Markov blankets within Markov blankets, extending from cellular organelles and DNA to elements within the environment.

The Hierarchically Mechanistic Mind (HMM) provides a more in-depth explanation for how this hierarchy is instantiated. It is a neurobiological framework developed specifically to explain the structure of the brain through a synthesis of FEP and major psychological paradigms to link small-scale neural behaviors all the way to human-specific phenotypes, behaviors and niches (Ramstead et al. 2018). It conceives of the brain as

> an embodied, complex adaptive system that adaptively minimises the entropy of its internal (i.e., sensory and physiological) states through recursive interactions between hierarchically organized, functionally differentiated neural subsystems. This hierarchy ranges from lower-order, highly segregated neurocognitive systems responsible for sensorimotor processing, through to the highly integrated association areas that underlie the sophisticated cognitive faculties unique to humans. The HMM resonates with structural and functional imaging studies in network neuroscience, which show that the brain entails a multiscale hierarchical

> organisation characterised by the repeated encapsulation of smaller neural elements in larger ones. Predictive coding approaches suggest that this sort of architecture entails a hierarchical generative model that minimises prediction error via recurrent message-passing between cortical levels, affording a neurobiologically plausible, mechanistic theory of the functional integration of anatomically segregated neural networks. (p. 9)

Thus, the foundation of the HMM lies in the assertion that the human brain is structured hierarchically, composed of interacting neurocognitive mechanisms that operate dynamically and reciprocally (Ramstead et al. 2018). According to this perspective, the lower or peripheral levels of the cortical hierarchy encompass distinct and specialized neural mechanisms dedicated to sensorimotor processing, known as "domain-specific" systems. Conversely, the higher or deeper layers of the hierarchy are characterized by developmentally flexible and extensively interconnected mechanisms, termed "domain-general" systems.

Two key distinctions are apparent here. Firstly, the neural hierarchy can be thought of as a fractal or nested modular hierarchy, characterized by the repetitive encapsulation of smaller neuronal elements within larger ones across various spatial, temporal, topological, and functional neural scales—essentially, a system of Markov blankets within Markov blankets. Secondly, a neurocognitive mechanism is defined as a neural subsystem operating across any spatiotemporal scale, spanning from individual neuronal populations to macroscopic brain regions. These mechanisms entail a dynamic, bidirectional relationship: they exhibit specialized functional processing facilitated by dense, short-range connections intrinsic to their scale (local integration), while also achieving global (functional) integration with other neural subsystems through relatively sparse, long-range connections (such as extrinsic cortico-cortical connections) (Park & Friston 2013).

Consequently, the HMM suggests a complementary interplay between functional segregation and integration: all neurocognitive mechanisms comprise a subset of cells with shared specialized functions, yet they are functionally integrated due to their distal connections with other subsystems. What’s more, the HMM fits naturally with and elaborates on the ideas within the FEP framework – the minimization of free energy is the mechanism that allows for the specialization and integration of neurons and why the structure of the brain is hierarchical in the sense of bottom-up and top-down (also, specific and general) dynamics guiding cognition.

IV.8. FEP and Mental Representation

As has been discussed, the sensorimotor loops proposed by the FEP framework work at the core of living systems to produce cognition and are co-determined by the inner and outer states of Markov blanketed system (Friston 2011; Clark 2016). Such a living system makes sense of its environment in an “organism-relative” way, making use of its own body (and mind, if it has one) to minimize the prediction error and free energy of its internal states through sensation and action, keeping it in a non-steady state thermodynamic equilibrium (Friston 2010). While this seems like a very distant starting point for describing social phenomena, it has the advantage of offering an integrated view of biological systems from cell to self and provides the groundwork for extending that view into social ontology.

The FEP framework first and foremost, provides a way to formalize the physics behind living systems, showing that self-creation, inference, and self-maintenance are all part of the same process; that is, the variational nature of free energy describes how these seemingly different processes are guided by the same value (Friston 2010; Friston, Sengupta, & Auletta 2014). By becoming statistically separated from their environment (forming a Markov Blanket), living systems at once infer the causes of their environment, their own states, and keep the dynamically coupled interaction between environment and living system in tact. The environment shapes the organism and the organism in turn shapes its environment. The only stable point within their dynamics becomes the interaction itself, as the states of both the environment and the living system are constantly dynamically changing. By treating interaction itself as the building block of living systems, it enables the nested, MB structure described by Ramstead and colleagues (Ramstead et al. 2018; Ramstead et al. 2019; Badcock, Friston, & Ramstead 2019) to unify cellular, organismic, social, and ecological dynamics. As the authors highlight, this provides potentially valuable avenues for future interdisciplinary research, allowing for the integration of a variety of perspectives from the social and hard sciences (Friston 2010; Ramstead et al. 2019).

As discussed previously, the current concept of mental representation – used in Searle and Guala's ontologies – is traditional, brain-based, contentful, and feed-forward and has been challenged by enactivist and interactionist philosophy (Varela, Thompson, & Rosch 1991; Bickhard & Terveen 1994; Bickhard 2008; Hutto & Myin 2013). The approach to representations within and without the enactivist framework emphasizes the role of body and environment in cognition, but is extremely diverse (Shapiro 2011; Orlandi 2016; Newen, De Bruin, & Gallagher 2018) with many claiming that mental representation is entirely unnecessary for explaining human and other living things' cognition or behavior (Pfeifer & Bongard 2006; Chemero 2009; Di Paolo & Thompson 2014; Hutto & Myin 2013). Others claim that some form of content-bearing mental representation is necessary for explaining complex behavior, even if embodied and extended to the environment (Prinz 2009; Clark 2022).

Self-creation, inference, and self-maintenance being all part and parcel of the same energy-minimization process means that the act of representation necessarily entails existing, inferring, and interacting. Organisms "represent" by virtue of being interactive models with their environments that reflect the environment's causal regularities. In this way, the FEP gives us a logic that can account of the four issues outlined in Searle and Guala. The individual and social are tied together, since the two mutually determine each other and aid each other in allostasis: norms are both prescriptive for behavior and descriptive of the institutional expectations, as both are part of energy minimization for the social-individual system. The physical and mental are connected, since inference or "cognition" happens already on a cellular level, instantiated by the physical interactions between organism and environment. Animal and human cognition do not have to be separate: single-celled bacteria, ferns, and human beings are all minimizing free-energy. This FE minimization can of course come in many shapes and forms and in the case of human beings, creates and maintains a complex and varied cultural niche. And lastly, there is no need to view social reality as either top-down or bottom-up. Just as the dynamic, interactive process between organism and environment is co-regulated, social reality is also co-regulated. Individuals and institutions both influence each other and the process results in the whole system keeping its free energy at a minimum.

There is, however, some issues that need to be mentioned in regards to mental representation and the FEP. Most frameworks within the FEP literature claim compatibility with enactivist views, but are further divided into two camps. In Helmholzian (inferential and internal) and representational (Friston 2010; Friston, Sengupta, & Auletta 2014; Hohwy 2013; Clark 2016) conceptions, predictive coding is the logic underlying mental models:

> it offers a biologically plausible means, consistent […] with as much reliance on external scaffolding as possible, of internally encoding and deploying richly structured bodies of information. This is because each PP level (perhaps these correspond to cortical columns – this is an open question) treats activity at the level below as if it were sensory data, and learns compressed methods to predict those unfolding patterns. This results in a very natural extraction of nested structure in the causes of the input signal, as different levels are progressively exposed to different recodings, and re-re-codings of the original sensory information. These re-re-codings (I think of them as representational re-descriptions in much the sense of Karmiloff-Smith 1992) enable us, as agents, to lock us onto worldly causes that are ever more recondite, capturing regularities visible only in patterns spread far in space and time. Patterns such as weather fronts, persons, elections, marriages, promises, and soccer games. Such patterns are the stuff of which human lives, and human mental lives, are made. What locks the *agent* on to these familiar patterns is, however, the whole multi-level processing device (sometimes, it is the whole machine in action). That machine works (if PP is correct) because each level is driven to try to find a compressed way to predict activity at the level below, all the way out to the sensory peripheries. These nested compressions, discovered and annealed in the furnace of action, are what I (following Hinton 1990) would like to call "internal representations". (Clark 2015, p. 5)

Such representation is very much content-bearing and internal, but based on a nested Markov blanket structure and very much embodied and extended to the environment. It is, after all, based on the sensorimotor loops described by the FEP, grounding the organism in a dynamic co-determination with its environment.

The debates around whether content-bearing representations are necessary or useful for explaining cognitive processes are still unresolved, with many open questions (Van Es & Myin 2020). Gładziejewski (2015) has made a detailed argument for keeping predictive processing (and thus FEP) representational. Conceiving of representations through the FEP framework makes them structural-functional. In essence, this means that FEP representation is not banking only on some neural structures standing in for phenomena in the environment, but that the way an agent interacts with the environment (the functional use of the representation) can be added to remedy the problems with the structural account of representations. On top of that, he has posited that because of this, a predictive processing (FEP) conception of representation withstands the enactivist critiques discussed in the previous sections.

Van Es and Myin (2020) have in turn argued, that the representational outline put forth by Gładziejewski does not actually present a solid case for why representations should be reified:

> The move to include function, or use, in the definition of representation is *prima facie* convincing: if a representation is only the type of relation that enables usage, it has to have causal effect. There are now, however, two ways in which 'use' is employed in representations. This invites a circle that is difficult to escape from. A representation is defined by its *use*. Only when a composition of driftwood is *used* as a map, can the composition count as a representation for the structure of the terrain it is used for (Ramsey 2007). However, as representations figure in an explanatory context, the *usefulness* of a representation, in turn, is defined in terms of the representational relation it bears onto the target. In other words, the way the relations are used are what make them representational, and their representational status is what makes them useful. (p. 5)

The problem, according to Van Es and Myin, is that there is a circularity within Gładziejewski's argument. A structural similarity is used functionally, because of its representational relation to the world, while a structural similarity is a representational relation to the world, due to the way it is used. The circularity does not justify the causal necessity of a representation.

The debates around whether the FEP should be interpreted representationally or not has led to its diverse applications to different emergent phenomena (e.g. vision, memory, cellular biophysics, sociocultural structures, among others), mixing both Helmholtzian and Gibsonian approaches. It has also been leveraged to generate a special brand of interactionist theories by combining it with an ecological perspective to cognition (Gibson 1979**;** Rietveld & Kiverstein 2014; Kiverstein 2018; Van Es & Myin 2020). Such a conception emphasizes dynamical patterns that the organism-environment coupled system tends toward, with the environment playing a crucial role not just in enabling cognition, but participating in it.

The framework outlined for a variational sociocultural ontology discussed previously (Fabry 2018; Veissière et al. 2019; Ramstead et al. 2019; Ramstead, Friston, & Hipólito 2020) lies ambiguously somewhere between the Helmholzian and Gibsonian camp. Ramstead et al.

(2019) claim that they posit a multiscale integrationist formulation of the boundaries of cognition, transcending the internal-external dichotomy. This framework is meant to be a pluralist ontology and methodology, treating "as a system sets of states that evince a robust enough form of conditional independence" (p. 543). This conditional independence is determined by the systems in question, through the FEP.

> This ontology implies, that any given cognitive system has a plurality of boundaries relevant to their scientific study; namely, the boundaries of its relevant subsystems. Our claim is that which among these are the *most relevant* will depend on the phenomenon being studied and the explanatory interests of researchers. Some of these boundaries are internal to the systems—these are boundaries of relevant *subsystems* nested in the whole system or organism (e.g., cells, ensemble of cells, organs, etc.); other boundaries separate the organism from its external environment (e.g., the skin membrane); and others still extend outwards to include the organism and external, worldly states (e.g., constructed niches and patterned cultural practices). (p. 543)

An ontology which is dependent on relevant explanatory interests of researchers using it, is in effect an epistemology – it is fundamentally a way to know about the world in order to answer a specific question. It is also bold to claim that a framework built using mathematical formalism is truly an ontology – it descriptive. So, it is unclear what exactly Ramstead and colleagues mean by such a pluralistic ontology or what it means for mental representation. Either way, an FEP theoretical basis for cognition offers a novel view of representation as at least extended to the body and environment – rather than being squarely in the brain – formed through sensorimotor loops, or potentially not necessary at all (Chalmers 2009; Van Es & Myin 2020; Clark 2016; 2022).

This novel "ontological" framework allows for both incentives and a sort of deontic power. For incentives, the system attempts to minimize error and this can be translated into optimal control and decision:

> The free-energy formulation starts with a free-energy bound on the value of states, which is specified by priors on the motion of hidden environmental states. These priors can incorporate any cost function to ensure that costly states are avoided. States with minimum cost can be set (by learning or evolution) in terms of prior expectations about motion and the attractors that ensue. In this view, the problem of finding sparse rewards in the environment is nature's solution to the problem of how to minimize the entropy (average surprise or free energy) of an

> agent's states: by ensuring they occupy a small set of attracting (that is, rewarding) states. (Friston 2010, p. 135)

On the other hand, the generation of predictions and their constant updating through prediction error ensures that novel affordances arise from repeated interactions with the environment. This can be thought of as ways in which new norms and cultural "content" can arise. Such content is "built on a foundation of contentless basic mental processes that acquire content through immersive participation of the agent in social practices that regulate joint attention and shared intentionality" (Ramstead, Veissière, & Kirmayer 2016, p. 1)

This is an interesting, but potentially confusing, combination of both contentless and content-driven representation. According to Ramstead and colleagues (2016), this dual framework allows for brain-based mental representations, while also treating the body as part of the sense-making of affordances in the environment. This

> integrative framework we propose bridges cognitive and social sciences to provide (i) an expanded concept of 'affordance' that extends to sociocultural forms of life, and (ii) a multilevel account of the socioculturally scaffolded forms of affordance learning and the transmission of affordances in patterned sociocultural practices. The context of the present discussion is the search for the 'natural origins of content' (Hutto and Satne 2015). We hope to contribute to the naturalistic account of the emergence of semantic content, that is, of the evolution (in phylogeny) and acquisition (in ontogeny) of representational or propositional content. (p. 2)

The FEP is used once again to bridge this gap (while also delineating a border). As discussed previously, it is still intensely debated whether preconceptual, embodied cognition can suddenly create content-driven mental representation.

From the point of view of contributing a potential new ontology, free energy-based accounts at least explain why institutions exist – namely, for the constraint and enabling of certain cooperative strategies, which stabilize sociocultural coordination. Because the FEP is fundamentally a process of interaction-stabilization and a delineation of self and environment, the content of self and environment is determined by the process itself. Thus, it better accounts for why institutions matter (they take part in this process of interaction and co-determination) and also provides a way for explaining how the deontic power of institutions doesn't come from arbitrary, collective rule-recognition and manipulation – as Searle describes – but from the dynamic shifting of the conditions that allows for the social-individual interaction to persist. This allows for both a gradual or sudden change in the conditions of the interaction, and thus sudden or gradual institutional change.

For example, as investors put money into stocks, those stocks in return change their value, which then changes the behavior of the investors. These kinds of feedback loops exist locally (such as the interaction between stock price and investment behavior), but are also always embedded in larger, more distal interactions (such as the impact of foreign economies' performance, natural disasters, worker strikes, political outcomes, etc.). It is of course possible to break down the complexity into its parts, but as discussed previously, due to the nonlinear nature of sociocultural phenomena, the separate analysis and reassembly of a system's constituent parts does not describe its emergent properties or provide much predictive power (for a detailed mathematical formalism of how emergent properties have causal influence on their constituent parts and the importance of synergy, see Mediano et al. 2022).

Another example of an institution that can be better described through the lens of free energy minimization is, as mentioned in the introduction, a favorite topic of social ontologists: traffic rules. Pedestrian crossings may seem like a straightforward endeavor with clear rules and behaviors, but the reality is often a lot more fluid and dynamic. Take street crossings in Vietnam, for instance: while technically, pedestrians should have right of way when crossing streets the same as they do in most other countries, in Vietnam the interaction between drivers/motorists and pedestrians is much more fine-tuned. Drivers do not generally wait for pedestrians to cross – instead, pedestrians walk at a steady pace across crosswalks and drivers pay careful attention to avoid them as they pass by. When looked at through the lens of FEP, we can understand that the interaction is a result of careful sensorimotor coordination which describes the dynamic and fluid nature of the crossing, while also being able to explain why this specific execution exists: the slowing of the heavy traffic flow in Vietnamese cities is significantly minimized, while also allowing pedestrians to cross.

IV.9. Interlude on a Potential Social Ontology Based on the FEP

This dynamic process is very different from the idea of "things" on the outside being represented in the mind, then driving action, creating a linear description for the brain. The FEP paints a picture of the environment and agent co-creating cognition, which is fundamentally an interactive process: the agent and its environment act on each other and are thus dynamically transforming constantly. This is not how the social ontologies of Searle and Guala treat cognition, where individual minds represent the external world and then act on it – where cognition is understood through traditional representation, which acts as the building block of shared beliefs, stable rules, or Schelling focal points. The FEP instead provides a cognitive framework that conceives of representation as ever-changing, constructed, and distributed over brain-body-environment. The framework thus ties together the biophysical levels of a living system, up to cognition and social cognition, all the way to social reality. Thus, the HMM claims to lay the groundwork for an ambitious integration of seemingly distinct fields. Using

> comparative, cross-cultural, computational, and dynamical approaches stemming from evolutionary psychology also allow[s] us to study in great detail the (epi)genetic mechanisms responsible for the emergence, transmission, and acquisition of our species-typical adaptive priors. Finally, computational models and simulation studies enable us to model how different levels of dynamical activity interact, allowing neuroscientists to explore how the biobehavioural phenomena described and studied by psychologists reflect adaptive free-energy minimisation within and across different spatiotemporal scales. The outcomes of such analyses can then be confirmed through real-world observations and experimental work. (Badcock, Friston, & Ramstead 2019, p. 116)

In order to accomplish this ambitious goal of creating an ontology that spans from neuron to society, Ramstead, Badcock, and Friston expand on the HMM framework with Evolutionary Systems Theory (EST)[13] and introduce the ideas of variational neuroethology (VNE) and

[13] Evolutionary Systems Theory (EST) is an interdisciplinary framework that combines ideas from evolutionary biology and systems theory to explain how complex, self-organizing systems – ranging from biological organisms to ecosystems, societies, and even technologies – evolve and adapt over time (Jantsch 1980; Laszlo 1987; Capra 1996). Some key elements: systems (whether biological, social, or ecological) are dynamic, constantly changing over time due to interactions among their components and external influences; systems are hierarchical as living organisms and other systems are made up of nested, hierarchically structured components, constantly interacting with each other. Additionally, it asserts that systems evolve through adaptation to their environments and co-evolution with other systems. This means that changes in one system can induce changes in others, leading to a co-evolutionary dynamic. EST also highlights that evolutionary processes are nonlinear and often involve

variational ecology (VE) (Ramstead et al. 2018; Badcock, Friston, & Ramstead 2019; Ramstead et al. 2019).

Variational neuroethology integrates different levels of organization and cognition through a multiscale FEP combined with EST to give us a clearer picture of how single cells to entire groups of individuals can organize around minimizing free energy. As outlined above, living systems are dynamical systems coupled to their environment and are thus subject to constant change in general selection dynamics – usually termed Darwinian natural selection for biological systems – which entails variation, selection, and retention. EST is concerned mainly with how such varying systems change and evolve over time. It lends itself well to explain the behavior of self-organizing systems (such as the brain, the biosphere, or social systems), where the system autonomously produces and selects locally interacting components within it to replicate or improve.

Ramstead and colleagues (2018) propose that FEP can be combined with EST: the minimization of free energy is the mechanism behind EST, and can be combined with Tinbergen's four levels of inquiry: adaptation, phylogeny, ontogeny, and mechanism. There are different dynamics within these four levels. The processes that generate 'functional adaptations' within species to the environment over evolutionary time, are adaptations through natural selection. They also encompass intergenerational 'phylogenetic' mechanisms that induce evolutionary change by instigating heritable differences between groups, such as epigenetic and cultural inheritance. Additionally, the theory addresses 'ontogenetic' processes that unfold over an individual's lifetime, such as gene-environment interactions. Finally, it accounts for the proximate 'mechanisms' driving psychology and behavior in real-time, characterized as ecobiopsychosocial dynamics. This is important due to the explanatory constraints for biological systems described by Ao (2005): laws governing the structure of evolutionary dynamics (e.g. natural selection) and laws governing the structure of the dynamical components involved in those dynamics (e.g. genes). FEP mostly describes the general dynamics of change for adaptive

---

feedback loops, where the output of a system influences its future behavior, creating cycles of reinforcement or inhibition. Lastly, EST posits that selection processes occur at multiple levels – from genes to individuals to groups – leading to the evolution of complex traits that are shaped by interactions within and across these levels.

active inferential systems (Markov blanketed systems minimizing their free energy), and thus does not give many details for the specific components in those dynamics. This is why it must be combined with EST, to provide specifics for what these general dynamics look like (e.g. a human being, living and creating a cultural niche).

VNE gives a formal account of how nested Markov blankets (MBs) can self-organize into semi-stable, multiscale dynamical systems (Ramstead et al. 2018; Ramstead et al. 2019).

> This ontology extends the MB formalism as a method of individuating systems under the FEP. As discussed above, a MB is a set of states that separates a system (a set of *internal* states) from *external*, non-systemic ones, which operationalises the idea that systems only exists *per se* if they are endowed with a robust form of conditional independence. According to VNE, the dynamics of living systems can be cast as active inference over *recursively nested* MBs. Indeed, the MB ontology can be reiterated recursively, such that the MBs at any one scale are composed in turn of MBs at the scale above and below – which are also made of MBs, and so on, all the way up, and all the way down. (Ramstead et al. 2019, p. 192)

Each MB is considered as a distinct entity (such as a particle), and its dynamics are summarized by combinations of blanket states that change gradually. This structure also aligns with order paramenters of slow eigenmodes described in synergetics and solutions on the center manifold (Haken 1983; Davis 2006).

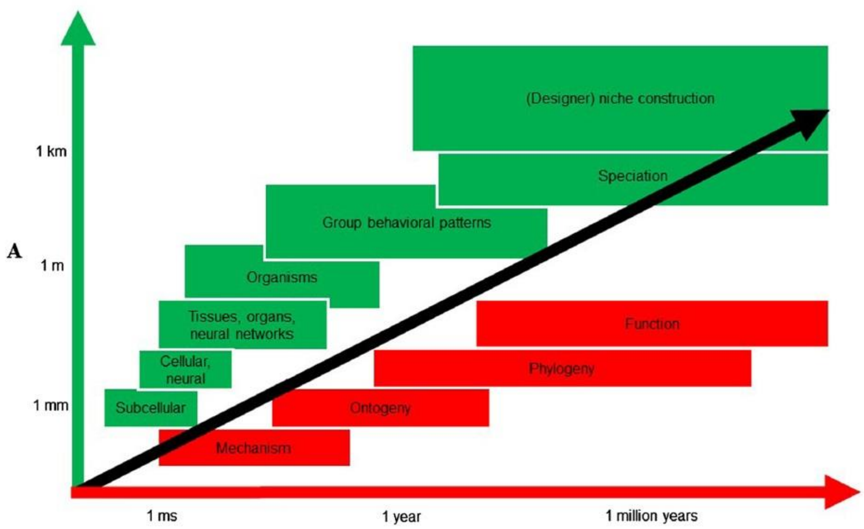


Fig. 2. *Explanatory scope of variational approach*. Variational neuroethology leverages the FEP to explain the adaptive free energy bounding dynamics of living systems across spatial and temporal scales. Here we indicate some of the scales at which these dynamics unfold. Given the intrinsic correlation between spatial and temporal scales, the phase space described here is populated mostly along its diagonal. Adapted from [92].

Figure 7

All in all, variational neuroethology describes how the different scales of interaction within an organism form a hierarchical dynamical system which is then used adaptively across different spatiotemporal scales (Ramstead et al. 2018; Figure 7, Ramstead et al. 2019). In this multiscale framework, active inference is by definition a collective activity. The entire ensemble of nested Markov blankets (MBs) is influenced, constrained, and governed by dynamics at higher scales, while the lower (microscopic) scales provide the (macroscopic) states at any given level. This arrangement demonstrates the circular causality that underpins synergetics, but it is extended to a recursive hierarchy of scales – essentially a hierarchical composition of blankets within blankets. Intuitively, the dynamics at one scale impose constraints (technically, create probability gradients) on the dynamics at other scales. Active inference reduces free energy gradients at each scale under the guidance or control of a generative model at the higher scale. This guidance is exerted through effects on sensory states, where circular causality implies that the action of any MB within an ensemble of MBs could be involved in sensing, action, or perception, depending on its role at the higher scale. An intuitive example is given by the authors:

> imagine you are an employee at an institution, where you transact your (microscopic) affairs with other personnel to self-evidence your prior beliefs that you are 'good at your job'. This would entail responding to corporate or institutional goals that emerge collectively (i.e., an implicit generative model at the macroscopic level). Note that your job may be homologous to an internal state at the institutional (macroscopic) level – relating only to other employees. Alternatively, you could be working on reception (i.e., a sensory state) or issuing press releases (i.e., an active state). (p. 193-194)

Variational ecology adds to this to better explain how the Markov blanket structure of a biological system and its immediate environment can be used to model and regulate the behavior of more distally related elements and relate them to free energy minimization (Ramstead et al. 2019). Using VNE to describe all of life and its mechanisms as being statistically discrete, neatly fitting into Markov blankets is an extremely ambitious framework (Bruineberg & Hesp 2018). While with a single organism and its immediate environment, the MB framework seems feasible, at larger scales systems' behavior may be too transient and not sufficiently robust. VE attempts to bridge this gap by building on the VNE framework's principles, adding the variational approach to niche construction (VANC) and the skilled intentionality framework (SIF) (Constant et al. 2018; Rietveld, Denys, & Van Westen 2018). VE claims that organisms aren't alone in minimizing free energy to regulate themselves and their interactions with their environments – individual organisms owe their robustness partly to their eco-niche's help in coordinating "dynamical patterns across larger spatiotemporal scales" and their environments in turn "learn" about their behavior (Ramstead et al. 2019, p. 189).

Niche construction is a fundamental part of how living beings engage with their environments, modifying their surroundings to maximize their chances of survival; it is also a process by which species steer their own and other species' evolutionary trajectory (Lewontin 1982; Odling-Smee et al. 2003; Laland et al. 2016). As Ramstead and colleagues (2019) put it:

> exogenetic resources of the niche such as desire paths can function as reliable indicators of surprise-and ambiguity-resolving actions, which, under the statistical conception of the phenotype entailed by the FEP, is crucial for maintaining the agent's continued existence. In short, the collective behaviour of an ensemble of agents provides a form of semiotics or a set of *possibilities for engagement with the niche* (a.k.a. *affordances*) that become relevant from the point of view of the needs and concerns of any single agent. (p. 196)

This is a significantly different way of describing organismic engagement with the environment than salience – it is a fundamentally different ontological grounding for sensing and acting on environments. Environmental salience implies that an environment inherently contains information, ready to be turned into representations by a biological agent, whereas affordances describe possibilities for interaction, which only gain meaning through the morphology and immediate/longterm allostatic needs of the agent (Rietveld & Kiverstein 2014; Ramstead et al. 2016).

Thus, the "field of affordances" refers to the set of affordances that attract the organism at any given moment (Ramstead et al. 2019). This field is formed by the expected free energy gradients generated by the entire ensemble. When the organism engages with its niche, it performs active inference, resolving a local free energy landscape—a path of least action through the field of affordances. "Solicitations" are the specific affordances that effectively draw the organism into action–perception loops at a particular time. Therefore, the "landscape of affordances" encompasses all the affordances available in a niche at a specific moment. In this context, the landscape of affordances is the result of inference about "what would happen if I did that?" Affordance is thus an attribute of active, albeit hypothetical, engagement with the niche. Simultaneously, it reflects the learned (and hence experienced) world.

VANC takes advantage of expressing expected free energy as the sum of expected cost and ambiguity, suggesting that agents offload much of the computational work for expected free energy to their custom-built environments through the use of affordances (Ramstead et al. 2019). Niche construction can be seen as the collective activity where organisms modify their material environment to create clear, unambiguous structures (affordances) that can be used for active inference. Persistent changes to the niche demonstrate how environmental cues serve as clear indicators of action possibilities. These cues act as epistemic resources that highlight actions resolving the ambiguity of future observations while aligning with the organism's (prior) preferences embedded in its generative model.

The integrated view that emerges from combining VANC with the ontology of affordances can be summarized as follows. When you choose actions that offer the greatest affordances, you simultaneously learn about your niche. However, your niche includes other Markov blankets that must also be learning about you. These MBs can represent other conspecifics, cognitive tools, desire paths, etc. In essence, your interaction with the environment serves as sensory evidence for a niche (i.e. a landscape of affordances) that is attempting to model you, while the niche influences you through sensory impressions. From a higher perspective, observing your niche would show self-organization at a higher spatiotemporal scale, resembling a self-organizing, free energy minimizing process. This must occur if your niche and all its components achieve some form of ergodicity[14], as suggested by the Free Energy Principle. In other words, the network of conspecifics, their cognitive tools, and desire paths would look like the internal states surrounded by an MB, distinguishing your niche from another.

---

[14] Ergodicity refers to the property of a system where, over time, its behavior covers all possible states or configurations within its phase space. In other words, a system is ergodic if, given sufficient time, it will explore all accessible states according to their probabilities. This concept is important in statistical mechanics and dynamical systems, as it underpins the idea that time averages of a system's observables will converge to the ensemble averages. Ergodicity implies that the long-term behavior of a system can be understood by analyzing its statistical properties at a single point in time (Frigg 2008).

IV.10. Conclusion

In summary, the Free Energy Principle can be utilized for a variety of theoretical frameworks that explain the mechanisms and organization of living systems, extending across large distances and timescales (Friston 2010; Orlandi 2016; Clark 2016; Wiese & Metzinger 2017; Piekarski 2019; Ramstead et al. 2019). The FEP can be leveraged as a biophysical description of living systems' self-creation, self-maintenance, and of their allostatic regulation of environmental interactions. It describes all living things as a dynamical system which is statistically separate from their environments, separated by a Markov blanket. All such Markov blanketed living systems are fundamentally involved in resisting entropy, inferring the states of their environment, bounding the entropy within their internal states, and acting on their environments all according to a free energy limit. This process is ultimately cast as variational Bayesian inference, which implies that all living systems are models of their environments, aiming to maximize the truth value of their predictions, thus keeping their states and the states of the environment predictable.

The HMM then takes as its basis the FEP and its implications about living systems to describe how the brain instantiates a hierarchical, Bayesian structure. In essence, the brain is structured as a set of hierarchically nested MBs, making it an embodied, complex system that minimizes the entropy of its internal states through recursive interactions between hierarchically organized, functionally differentiated and integrated neural subsystems. Not only does the mind perform free energy minimization through neural activity, but its neural organization and bodily connections embody the FEP (e.g. the amygdala regulates endocrine and autonomic responses in the body; in FEP language, these systems regulate each other through bottom-up and top-down signals to regulate the allostatic needs of the whole system (Barrett, Quigley, & Hamilton 2016). The amygdala is itself a "prediction" that expects sensory signals in the form of neurotransmitters ("prediction errors") from the endocrine system and the ANS.)

This point is expanded upon by variational neuroethology, which combines the HMM with Evolutionary Systems Theory and Tinbergen's four levels of inquiry. VNE thus illustrates that the logic of evolution can be cast as free energy minimization over extremely large spatiotemporal distances, resulting in unique adaptations, ontogenies, phylogenies, and mechanisms in different species. A whale's body, for instance, has been sculpted over millions of years to live in, traverse, and depend on oceans. Its body is sculpted to "predict" the dynamics of an ocean (e.g. streamlined shape and fins for fluid dynamics, specialized teeth for specific food sources, whale calls for social coordination, etc.), endowing it with species-typical tools, ontogenetic and phylogenetic development, and mechanisms that tend toward a free energy minimum in interaction with its specific environmental niche and allow for its survival.

Lastly, Variational Ecology takes VNE and the ontology of affordances and combine them to explain how distally related parts of living systems function as a whole. Such a whole is a Markov blanketed system which takes part in the prediction error minimization of all the Markov blanketed systems contained within it. It first posits that environments "afford" behavior possibilities for living things and that these affordances are dependent partly on the morphology of the living things.

> Grasping a glass, riding a bike, or improving an architectural design, for instance, can all be seen as a skilled individual's immediate responsiveness to affordances. An individual can respond to affordances thanks to abilities. In the relational approach to affordances developed in this chapter, the possession of the relevant ability is seen as necessary for being able to act on an affordance. (Rietveld, Denys, & Van Westen 2018, p. 2)

It also describes how organisms acting on their environments induce changes in the environments that can also be described as the environment adjusting its states to better reflect the behaviors of living things within it. When animals walk through a forest, they find paths that are most optimal for conserving energy, providing camouflage, and have a safe and stable ground to walk on. Given repeated traversals, vegetation diminishes in these areas and trails form – these trails can be thought of as evidence of animals predicting which aspects of their environments afford the most optimal possibilities for traversal *or* as the environment predicting animals' behavior and adjusting its states to better model these behaviors.

The common theme in these frameworks is not just prediction error minimization (free energy minimization). They are also fundamentally interaction-based frameworks. Bickhard (2008; 2009) along with Hasson & Frith (2016) and Cisek (1999), have put forth similar frameworks for cognition. The common theme within these perspectives is that an ontology of the social must take as a conceptual basis the interaction between systems itself, rather than the elements of the interaction. As discussed previously, this is a departure from individualist frameworks which posit that the sum of individuals' behaviors or attitudes can be cast as the basic units of social life, which continues to be a theme in social ontology. The debate between more behavior-oriented (like Guala's equilibria) and belief-oriented (Searle's constitutive rules) can be seen then as accentuating either half of an interaction – the bottom-up, incentivized, behavioral aspect or the top-down, normative, attitudinal aspects of social interaction. Interaction-based approaches instead take interactions as the basic unit of social ontology. While seemingly dodging an important question and instead embracing paradoxical thinking – how can top-down, normative processes drive behavior when bottom-up, incentivized processes guide it instead (or vice versa) – interactionism actually provides a novel perspective to explain the often paradoxical nature of sociality by adding insights from the Free Energy Principal framework.

V. Interdisciplinarity

V.1. Statement on Interdisciplinarity

This thesis takes as its central question how the conceptualization of mental representations influences theory in social ontology. While based primarily in philosophy, the thesis incorporates a wide variety of perspectives from diverse academic disciplines in order to investigate this question. First, it explores philosophical and economics perspectives from John Searle (2010) and Francisco Guala (2016) to frame the current debate in social ontology, which centers around whether institutions are normative or incentivized. It then incorporates perspectives from anthropology and social and cognitive psychology to elucidate how current conceptualizations of mental representations impact theory in social ontology (Tomsello 1999; Bickard 2008; Azaad, Knoblich & Sebanz 2021). Exploring the cognitive mechanisms behind mental representation and action, the thesis then adds perspectives from dynamics (May 1976; Busye 2017; Friston 2010; Ramstead, Veissière, & Kirmayer 2016; Ramstead et al. 2019; Badcock, Friston, & Ramstead 2019) cognitive neuroscience, biophysics, and statistical modeling to discuss novel ideas about cognition, representation, and self-organization in living systems. Finally, the thesis will integrate these perspectives to offer ways forward for research in social ontology and argue for a reframing of mental representations to transcend current debates and bridge theoretical problems.

VI. Conclusion

VI.1. The Free Energy Principle, Social Interactionism, and Mental Representations

The novel insights about cognition expressed by the FEP framework are compelling as explanations for sensation, thinking, and action and show promise to shed light on interpersonal and societal dynamics as well (Friston 2010; Ramstead et al. 2019; Knoblich, Azaad, & Sebanz 2021). They reconceptualize mental modeling as well, moving away from feed-forward, computational, encodingist representation to sensorimotor loops, Bayesian prediction, and autopoiesis (Clark 2015; 2016). It is telling that in the variational free energy accounts of the social and ecological, there is an increased focus on dynamics and statistical modeling. It is, however, worth investigating why this stark difference between a dynamic, variational account of social ontology and more traditional accounts based on rules and equilibria rests on a specific brand of mental representation.

Whether the FEP can work as a satisfactory account of biophysics and autopoiesis, neural dynamics, social cognition, environmental-agent interaction, cultural dynamics, emergence, and consciousness is an open question. Such a unifying theory is extremely ambitious and runs the risk of having almost no predictive power or verifiability (Friston et al. 2023). This expansive scope is enabled by the nested, hierarchical structure of Markov blankets which together create a system which is autonomous, homeostatic, and adaptive – its reactions to the environment demonstrate a generative model operating on several levels, with temporal depth and flexibility in possible responses for maintaining its states (Kirchhoff et al. 2017). In other words, the seemingly endless number of emergent phenomena that can be folded into the FEP ontological conception is a demonstration of exactly why it is difficult to delineate *social* ontology from cognition or biophysics. The ambitious scope of the FEP instead focuses on an interconnected, multi-layered decision-making system, whether it’s social groups, single cells, or ecosystems performing adaptive allostasis.

The FEP framework in this way provides a new approach to social ontology: there is considerable difficulty in delineating where emergent phenomena end and where they begin.[15] The MBs of a free energy-based ontology are "intradependent" – they not only depend on each other, but paradoxically define and redefine their own borders (Kirchoff et al. 2017; Palacios et al. 2019). This is a crucial ontological difference from the work of Searle (2010) and Guala (2016), both of whom treat individuals as firmly bounded, stable entities. What's more, they speak of mental representations in the same fashion, allowing for mental representations to do the hard work of defining where the cognition ends, and action/sensation/environment begin.

The question of representations is really then the question of where we draw the line between cognition (or mind) and everything else. Is there such a line? This question demands a serious reckoning with and connects to questioning the goals of social ontology. Is it to find an ontology that can describe all social systems? Is it to model the social for other purposes? What debates about representation and the constitutive debate about where cognition ends and everything else begins are echoed by what the FEP description of life describes: the meaning and nature of things depends largely on the interaction that describes them. If we model social systems according to a certain model, and find that our ontology is correct, we will act on the world according to that ontology until it changes and our ontology is no longer true. If we instead try and determine a "stable" ontology, we will find ourselves permanently bogged down trying to draw theoretical boundaries around systems which are fundamentally interconnected, co-dependent and co-constitutive.

I have argued that the way Searle and Guala use mental representation in their work is problematic for four reasons. This representational ground is encodingist (brains encode outside

---

[15] This is one of the FEP-based ontology's biggest points of weakness, since it is unclear how an internalist, representational FEP can be leveraged to describe an emergent sociocultural level – if agent and environment both predict each other, are they both engaging in representational "cognition"? Radical Extended Cognitive Neuroscience (RECN) argues that the extendedness of cognitive processes are evident when the variables describing one system also act as the parameters influencing changes in another system, and vice versa. In such cases, treating the agent and its environment as separate systems is merely a matter of explanatory convenience. In reality, the dynamics of both systems are so closely correlated and integrated that they are better understood as constituting a single, unified brain–body–world system (Favela 2024). To treat representation as strongly internal is an explanatory tool, not necessarily a good ontological commitment.

stimuli into representations within the brain), feed-forward (representations are created by "copying" an agent's environment in an attempt to have an inner version of "true" reality), brain-centric (it is primarily the brain which deals with forming coherent, cognitively useful representations), and internalist (representations exist within the agent and are not co-constituted by the environment or other agents). The problematic assumption here is that mental representations are so stable. Without this imagined stability, the theoretical work that follows – whether more based in cooperation problems and economic games or in constitutive rules and deontology – becomes moot.

My examination of the Free Energy Principle in cognition and its application to social phenomena was primarily to challenge the assumption which treats mental phenomena in the above fashion. I am not advocating for the reappraisal of different theoretical frameworks to all fit neatly into the conceptual framework of free energy minimization. Whether this framework continues to be supported by evidence from neuroscience and biophysics remains an open question, but it is one of the only formal models that has attempted to grapple with the challenges put forth by enactivist and interactionist philosophers to currently accepted conceptions of cognition. Additionally, the expansive literature which makes use of the FEP – of which only a fraction was covered in this thesis – still conceives of cognition as fundamentally representational (Clark 2016, 2022). It is simply that free energy-based representations are not stable in the same manner as traditional, encodingist ones: they are Bayesian approximations that are fundamentally action-oriented. The input and output of the agent is integral to representation, since the agent itself is a model of its environment. Thus, FEP representations have fuzzy outlines and include the active part of cognition, not only its perceptive aspect.

The conceptual shift from encodingist cognition to an interactionist one is not important for social ontology for its integrative and predictive power. Friston (2010) himself admits as much: "What does the free-energy principle portend for the future? If its main contribution is to integrate established theories, then the answer is probably 'not a lot'" (p. 136). It is instead to highlight and explore the interconnected, autopoietic, and nested nature of our cognition and its import on social life. Given such looping, self-defining systems, what questions should we ask about the social? Does it even make sense to ask about an *ontology* of the social, that can be separated from the society asking the question? After all, following the logic of self-creating systems, our conception of the social world shapes the social world and vice-versa. To isolate some inherent qualities of institutions, or social phenomena in general, may prove to be a thankless task. The interactionist view of cognition based on the FEP hints at a mind that is as much a creator of its social world as it is a receiver of it; that perhaps epistemology and ontology are more closely related than they seem.

References

Acemoglu, D., & Robinson, J. A. (2012). *Why nations fail: The origins of power, prosperity, and poverty*. Crown Publishers.

Akiyama, E., & Kaneko, K. (2000). Dynamical systems game theory and dynamics of games. *Physica D: Nonlinear Phenomena*, *147*(3–4), 221–258. https://doi.org/10.1016/S0167-2789(00)00157-3

Anderson, M. (2014). After phrenology: Neural reuse and the interactive brain (1st ed.). Cambridge, MA: The MIT Press.

Anderson, M. (2016). Précis of after phrenology: Neural reuse and the interactive brain. Behavioral and Brain Sciences. doi:10.1017/s0140525x15000631.

Ao, P. (2005). Laws in Darwinian evolutionary theory. *Physics of Life Reviews*, *2*(2), 117–156. https://doi.org/10.1016/j.plrev.2005.03.002

Aoki, M. (2001). *Toward a comparative institutional analysis*. MIT Press.

Ashby, W. R. (1947). Principles of the self-organizing dynamic system. *Journal of General Psychology*, *37*, 125–128. https://doi.org/10.1080/00221309.1947.9918144

Astle, D. E., Johnson, M. H., & Akarca, D. (2023). Toward computational neuroconstructivism: A framework for developmental systems neuroscience. Trends in Cognitive Sciences, 27(8), 726–744. https://doi.org/10.1016/j.tics.2023.04.009

Azaad, S., Knoblich, G., & Sebanz, N. (2021). Perception and Action in a Social Context. Cambridge: Cambridge University Press.

Badcock, P. B., Friston, K. J., & Ramstead, M. J. D. (2019). The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Physics of Life Reviews*, *31*, 104–121. https://doi.org/10.1016/j.plrev.2018.10.002

Barkow, J. H., Cosmides, L., & Tooby, J. (Eds.). (1992). *The adapted mind: Evolutionary psychology and the generation of culture.* Oxford University Press.

Barrett L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. Social cognitive and affective neuroscience, 12(1), 1–23. https://doi.org/10.1093/scan/nsw154

Barrett, L. F., Quigley, K. S., & Hamilton, P. (2016). An active inference theory of allostasis and interoception in depression. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1708). https://doi.org/10.1098/rstb.2016.0011

Barton, S. (1994) Chaos, self-organization, and psychology. *American Psychologist*, Vol 49(1), Jan 1994, 5-14

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms.* Cambridge University Press.

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

Bickhard, Mark H. (2009). The interactivist model. Synthese 166 (3):547 - 591.

Bickhard, M. H., & Terveen, L. (1995). *Foundational issues in artificial intelligence and cognitive science: Impasse and solution.* North-Holland/Elsevier Science Publishers.

Boyer, P., & Petersen, M. B. (2011). The naturalness of (many) social institutions: Evolved cognition as their foundation. *Journal of Institutional Economics, 8*(1), 1–25. https://doi.org/10.1017/S1744137411000300

Bratman, M. E. (1993). *Shared intention*. In A. Meijers (Ed.), *Joint action: Proceedings of the 17th International Conference on Artificial Intelligence* (pp. 4-15).

Bratman, M. E. (2014). *Shared agency: A planning theory of acting together*. Oxford University Press.

Brinton, M. C., & Nee, V. (Eds.). (1998). *The new institutionalism in sociology*. Stanford University Press.

Britannica, T. Editors of Encyclopaedia (2024, October 10). naturalism. Encyclopedia

Britannica. https://www.britannica.com/topic/naturalism-philosophy

Brown, E. C., & Brüne, M. (2012). The role of prediction in social neuroscience. Frontiers in human neuroscience, 6, 147. https://doi.org/10.3389/fnhum.2012.00147

Bruineberg, J., & Hesp, C. (2018). Beyond blanket terms: Challenges for the explanatory value of variational (neuro-)ethology: Comment on “Answering Schrödinger’s question: A free-energy formulation” by Maxwell James Désormeau Ramstead et al. *Physics of Life Reviews*, *24*, 37–39. https://doi.org/10.1016/j.plrev.2017.11.015

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2018). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, *195*(6), 2417–2444. https://doi.org/10.1007/s11229-016-1239-1

Capra, F. (1996). The Web of Life: A New Scientific Understanding of Living Systems. Anchor Books.

Chemero, A. (2009). *Radical embodied cognitive science.* MIT Press.

Cisek, P. (1999). Beyond the computer metaphor: Behaviour as interaction. Journal of Consciousness Studies 6 (11-12):11-12.

Clark, A. (2015). Predicting Peace The End of the Representation Wars-A Reply to Michael Madary. *Open MIND*, *7*(R), 1–7. https://doi.org/10.15502/9783958570979

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press. https://books.google.at/books?id=Yoh2CgAAQBAJ

Clark, A. (2017). How to Knit Your Own Markov Blanket: Resisting the Second Law with Metamorphic Minds. *Philosophy and Predictive Coding*, *Rupert 2009*, 1–19. https://doi.org/10.15502/9783958573031

Clark, A. (2022). Extending the Predictive Mind. Australasian Journal of Philosophy. 102. 1-12. 10.1080/00048402.2022.2122523.

Constant, A., Ramstead, M. J. D., Veissière, S. P. L., Campbell, J. O., & Friston, K. J. (2018). A variational approach to niche construction. *Journal of the Royal Society Interface*, *15*(141), 20170685. https://doi.org/10.1098/rsif.2017.0685

Costall, A. (2006). 'Introspectionism' and the mythical origins of scientific psychology. *Consciousness and Cognition: An International Journal*, *15*(4), 634–654. https://doi.org/10.1016/j.concog.2006.09.008

Davis, M.J. (2006). Low-Dimensional Manifolds in Reaction−Diffusion Equations. *J. Phys. Chem. A* 2006, 110, 16, 5235–5256 https://doi.org/10.1021/jp055592s

DiMaggio, P. (1998). The new institutionalisms: Avenues of collaboration. *Journal of Institutional and Theoretical Economics (JITE)*, 154(4), 696-705. Mohr Siebeck GmbH & Co. KG. https://www.jstor.org/stable/40752104

Di Paolo, E., & Thompson, E. (2014). The enactive approach. In L. Shapiro (Ed.), *The Routledge handbook of embodied cognition* (pp. 68–78). Routledge/Taylor & Francis Group.

Dobbin, F., Simmons, B., & Garrett, G. (2007). *The global diffusion of public policies: Social construction, coercion, competition, or learning?* Annual Review of Sociology, 33, 449–472. https://doi.org/10.1146/annurev.soc.33.090106.142507

Dołęga, K., Roelofs, L., & Schlicht, T. (2018). Introduction to Special Issue on “Enactivism, Representationalism, and Predictive Processing”. Philosophical Explorations 21 (2):179-186.

Engel, A. K., Friston K. J., & Kragic., D. (2016). 'Introduction: Where’s the Action?', in Andreas K. Engel, Karl J. Friston, and Danica Kragic (eds), *The Pragmatic Turn: Toward Action-Oriented Views in Cognitive Science.* Cambridge, MA; online edn, MIT Press Scholarship Online. https://doi.org/10.7551/mitpress/9780262034326.003.0001

Fabry, R. E. (2018). Betwixt and between: the enculturated predictive processing approach to cognition. *Synthese* 195 (6):2483-2518.

Farrell, H. (2018). *Shared challenges of institutional theories*. In *The Oxford Handbook of Political Institutions* (pp. 57-78). Oxford University Press.

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. Frontiers in Human Neuroscience, 4. https://doi.org/10.3389/fnhum.2010.00215

Fodor, J. A. (1975). The Language of Thought. Harvard University Press.

Fodor, J. A. (1981). Representations: philosophical essays on the foundations of cognitive science. Cambridge: MIT Press.

Fodor, J. A. (1983). The Modularity of Mind: An Essay on Faculty Psychology. Cambridge, MA: MIT Press.

Fodor, J. A. (1987). Psychosemantics: The Problem of Meaning in the Philosophy of Mind. MIT Press. Edited by Margaret A. Boden.

Frigg, R. (2008). A Field Guide to Recent Work on the Foundations of Statistical Mechanics.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K. (2011). Embodied inference: or "I think therefore I am, if I am what I think" In W. Tschacher & C. Bergomi (Eds.), *The implications of embodiment: Cognition and communication* (pp. 89–125). Imprint Academic.

Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 20130475. https://doi.org/10.1098/rsif.2013.0475

Friston, K. (2018). Does predictive coding have a future?. *Nat Neurosci* 21, 1019–1021 https://doi.org/10.1038/s41593-018-0200-7

Friston, K., Da Costa, L., Sajid, N., Heins, C., Ueltzhöffer, K., Pavliotis, G. A., & Parr, T. (2023). The free energy principle made simpler but not too simple. *Physics Reports*, *1024*, 1–29. https://doi.org/10.1016/j.physrep.2023.07.001

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. Biological cybernetics, 104(1-2), 137–160. https://doi.org/10.1007/s00422-011-0424-z

Friston, K., Sengupta, B., & Auletta, G. (2014). Cognitive dynamics: From attractors to active inference. *Proceedings of the IEEE*, *102*(4), 427–445. https://doi.org/10.1109/JPROC.2014.2306251

Fukuyama, F. (2011). *The Origins of Political Order: From Prehuman Times to the French Revolution*. Farrar, Straus and Giroux.

Gibson, J. J. (1979). The Ecological Approach to Visual Perception: Classic Edition. Houghton Mifflin.

Gilbert, M. (1987). *Modelling collective belief. Synthese*, 73(2), 185-204.

Gilbert, M. (2009). Shared intention and personal intentions. *Philosophical Studies* 144 (1):167-187.

Gilbert, M. (2014). The Nature of Agreements: A Solution to Some Puzzles about Claim-Rights and Joint Intention. In Manuel Vargas & Gideon Yaffe (eds.), *Rational and Social Agency: The Philosophy of Michael Bratman*. New York, NY: Oxford University Press. pp. 215-255.

Gilestro, G. F., Tononi, G., & Cirelli, C. (2009). Widespread changes in synaptic markers as a function of sleep and wakefulness in Drosophila. Science (New York, N.Y.), 324(5923), 109–112. https://doi.org/10.1126/science.1166673

Gładziejewski, P. (2015). Predictive coding and representationalism. *Synthese*, 193(2), 559–582.

Graeber, D., & Wengrow, D. (2021). *The dawn of everything: a new history of humanity*. First paperback edition. New York, Picador.

Gray, K. L. H., Barber, L., Murphy, J., & Cook, R. (2017). Social interaction contexts bias the perceived expressions of interactants. Emotion, 17(4), 567–71. https://doi.org/10.1037/emo0000257

Greenwood, R., Raynard, M., Kodeih, F., Micelotta, E. R., & Lounsbury, M. (2011). *Institutional complexity and organizational responses*. Academy of Management Annals, 5(1), 317-371. https://doi.org/10.5465/19416520.2011.590299

Greif, A., & Kingston, C. (2011). Political Economy of Institutions, Democracy and Voting. In *Political Economy of Institutions, Democracy and Voting* (Issue March). https://doi.org/10.1007/978-3-642-19519-8

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

Haken, H. (1983). Synergetics: An introduction. Non-equilibrium phase transition and self-organisation in physics, chemistry and biology. Springer Verlag, Berlin.

Haslanger, S. (2012). Resisting Reality: Social Construction and Social Critique. New York, US: Oxford University Press.

Hasson, U., & Frith, C. D. (2016). Mirroring and beyond: Coupled dynamics as a generalized framework for modelling social interactions. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 371, Issue 1693). Royal Society of London. https://doi.org/10.1098/rstb.2015.0366

Hawkins, J., & Blakeslee, S. (2004) *On intelligence.* Henry Holt and Company.

Hebb, D. O. (1949). *The organization of behavior; a neuropsychological theory.* Wiley.

Heintz, C. (2007). Institutions as Mechanisms of Cultural Evolution: Prospects of the Epidemiological Approach. *Biological Theory*, *2*(3), 244–249. https://doi.org/10.1162/biot.2007.2.3.244

Helmholtz, H. (1882). Thermodynamik der chemischen Vorgänge [Thermodynamics of chemical processes]. Leipzig: Barth.

Hindriks, F. (2003). Social Groups, Collective Intentionality, and Anti-Hegelian Skepticism. In: Sintonen, M., Ylikoski, P., Miller, K. (eds) Realism in Action. Synthese Library, vol 321. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-1046-7_14

Hilgetag, C. C., & Goulas A. (2020). 'Hierarchy' in the organization of brain networks. *Phil. Trans. R. Soc.* B375:20190319 http://doi.org/10.1098/rstb.2019.0319

Hodgson, G. M. (2006). What are institutions? *Journal of Economic Issues*, *40*(1), 1–25. https://doi.org/10.1080/00213624.2006.11506879

Hohwy, J. (2013). *The predictive mind.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199682737.001.0001

Hommel, B. (2015). The theory of event coding (TEC) as embodied-cognition framework. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.01318

Hutto, D. D. (2005). Knowing what? Radical versus conservative enactivism. Phenomenology and the Cognitive Sciences, 4 (4), 389-405.

Hutto, D. D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content.* MIT Press.

Hutto, D. D., Kirchhoff, M. D., & Myin, E. (2014). Extensive enactivism: why keep it all in? Frontiers in Human Neuroscience 8 (706):102178.

Illingworth, V. (1991). *The penguin dictionary of physics.* Penguin Books.

Jacob, P, "Intentionality", The Stanford Encyclopedia of Philosophy (Spring 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/spr2023/entries/intentionality/

Jantsch, E. (1980). The Self-Organizing Universe: Scientific and Human Implications of the Emerging Paradigm of Evolution. Pergamon Press.

Jarvie, I. C. (1986). Thinking about Society: Theory and Practice. Springer Verlag.

Jaynes, E. T. (1957). Information theory and statistical mechanics. Physical Review, 106(4), 620–630. https://doi.org/10.1103/PhysRev.106.620

Jordan J. S. (2003). Emergence of self and other in perception and action: an event-control approach. *Consciousness and cognition*, 12(4), 633–646. https://doi.org/10.1016/s1053-8100(03)00075-8

Jörg, T. (2011) *New thinking in complexity for the social sciences and humanities: a generative transdisciplinary approach*. New York: Springer ISBN 978-94-007-1302-4

Karmiloff-Smith, A. (1992). Beyond modularity: A developmental perspective on cognitive science. The MIT Press.

Karmiloff-Smith, A., Thomas, M.S.C., & Johnson, M.H. (2018). Thinking Developmentally from Constructivism to Neuroconstructivism: Selected Works of Annette Karmiloff-Smith (1st ed.). Routledge. https://doi.org/10.4324/9781315516691

Kauffman, S. A. (1993). The origins of order. Self-organization and selection in evolution. New York/Oxford: Oxford University Press.

Kauffman, S. A. (1995). At Home in the Universe: The Search for Laws of Self-organization and Complexity. Oxford University Press USA.

Kilner, J.M., Friston, K.J. & Frith, C.D. (2007). Predictive coding: an account of the mirror neuron system. *Cogn Process* 8, 159–166. https://doi.org/10.1007/s10339-007-0170-2

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. Journal of the Royal Society, Interface, 15(138), 20170792. https://doi.org/10.1098/rsif.2017.0792

Kiverstein, J. (2018). Extended Cognition. In: Newen, A., De Bruin, L. and Gallagher, S., Eds., *The Oxford Handbook of 4E Cognition*, Oxford University Press, New York, 19-40.

Klengel, T., Mehta, D., Anacker, C., Rex-Haffner, M., Pruessner, J. C., Pariante, C. M., Pace, T. W., Mercer, K. B., Mayberg, H. S., Bradley, B., Nemeroff, C. B., Holsboer, F., Heim, C. M., Ressler, K. J., Rein, T., & Binder, E. B. (2013). Allele-specific FKBP5 DNA demethylation mediates gene-childhood trauma interactions. *Nature neuroscience*, 16(1), 33–41. https://doi.org/10.1038/nn.3275

Laland, K., Matthews, B., Feldman, M. W. (2016). An introduction to niche construction theory. *Evol Ecol* 30:191–202.

Laszlo, E. (1987). Evolution: The Grand Synthesis. Shambhala Publications.

Levine, N. E. (2014). Marriage. obo in Anthropology. doi: 10.1093/obo/9780199766567-0016

Lewontin R. C. (1982) Organism and environment. In: Learning, development, and culture.

Mahoney, J., & Thelen, K. (2010). *Explaining institutional change: Ambiguity, agency, and power*. Cambridge University Press.

Manera, V., Schouten, B., Becchio, C., Bara, B. G., & Verfaillie, K. (2010). Inferring intentions from biological motion: A stimulus set of point-light communicative interactions. Behavior Research Methods, 42(1), 168–78. https://doi.org/10.3758/BRM.42.1.168

Maté, G., & Maté, D. (2022). The myth of normal: trauma, illness & healing in a toxic culture. New York, Avery, an imprint of Penguin Random House.

May, R. (1976). Simple Mathematical Models With Very Complicated Dynamics. Nature, 26, 457. https://doi.org/10.1038/261459a0.

Mediano, P. A. M., Rosas, F. E., Luppi, A. I., Jensen, H. J., Seth, A. K., Barrett, A. B., Carhart-Harris, R. L., & Bor, D. (2022). Greater than the parts: a review of the information decomposition approach to causal emergence. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 380(2227), 20210246. https://doi.org/10.1098/rsta.2021.0246

Merleau-Ponty, M. (1945). Phénoménologie de la perception . London: New York Routledge & K. Paul; Humanities Press.

Miles, L. K., Nind, L. K., & Macrae, C. N. (2009). The rhythm of rapport: Interpersonal synchrony and social perception. Journal of Experimental Social Psychology, 45(3), 585–9. https://doi.org/10.1016/j.jesp.2009.02.002

Morin, O. (2013). Three ways of misunderstanding the power of rules. *The Background of Social Reality: Selected Contributions from the Inaugural Meeting of ENSO*, 185–199. https://doi.org/10.1007/978-94-007-5600-7_11

Newen, A., Gallagher, S., & De Bruin, L. (2018). 4E Cognition. *The Oxford Handbook of 4E Cognition*, *October*, 2–16. https://doi.org/10.1093/oxfordhb/9780198735410.013.1

Noë, A. (2004). Action in Perception. MIT Press.

North, D. C. (1991). Institutions Published by : American Economic Association Institutions. *The Journal of Economic Perspectives*, *5*(1), 97–112.

Noy, L., Dekel, E., & Alon, U. (2011). The mirror game as a paradigm for studying the dynamics of two people improvising motion together. Proceedings of the National Academy of Sciences, 108(52), 20947–52. https://doi.org/10.1073/pnas.1108155108

Noyes, A., & Keil, F. C. (2020). Collective recognition and function in concepts of institutional social groups. *Journal of Experimental Psychology: General*, *149*(7), 1344–1359. https://doi.org/10.1037/xge0000711

O'Brien, G., & Opie, J. (2015). Intentionality Lite or Analog Content?: A Response to Hutto and Satne. Philosophia 43 (3):723-729.

Odling-Smee, J., Laland, K. N., Feldman, M. W. (2003). Niche construction: the neglected process in evolution. Princeton University Press.

Orlandi, N. (2016). Bayesian perception is ecological perception. *Philosophical Topics*, *44*(2), 327–351. https://doi.org/10.5840/philtopics201644226

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.

Ostrom, E. (2005), Understanding Institutional Diversity, Princeton, NJ: Princeton University Press.

Pacherie, E. (2005). Is collective intentionality really primitive?

Palacios, E. R., Isomura, T., Parr, T., & Friston, K. (2019). The emergence of synchrony in networks of mutually inferring neurons. Scientific reports, 9(1), 6412. https://doi.org/10.1038/s41598-019-42821-7

Park, H. J., & Friston, K. (2013). Structural and functional brain networks: From connections to cognition. *Science*, *342*(6158). https://doi.org/10.1126/science.1238411

Paus, T., Keshavan, M., & Giedd, J. N. (2008). Why do many psychiatric disorders emerge during adolescence?. Nature reviews. Neuroscience, 9(12), 947–957. https://doi.org/10.1038/nrn2513

Pfeifer, R. & Bongard, J. (2006). *How the body shapes the way we think: A new view of intelligence*. Cambridge, MA: MIT Press.

Piaget, J. (1970). *Science of education and the psychology of the child. Trans. D. Coltman.* Orion.

Piekarski, M. A. (2019). Normativity of predictions: A new research perspective. *Frontiers in Psychology*, *10*(JULY), 2010–2013. https://doi.org/10.3389/fpsyg.2019.01710

Pomiechowska, B., & Csibra, G. (2017). Motor activation during action perception depends on action interpretation. *Neuropsychologia*, *105,* 84–91. https://doi.org/10.1016/j.neuropsychologia.2017.01.032

Powell, W. W. (1991). *Neither market nor hierarchy: Network forms of organization*. In G. Thompson, J. Frances, R. Levacic, & J. Mitchell (Eds.), *Markets, hierarchies & networks: The coordination of social life* (pp. 265-276). Sage.

Powell, W. W., & DiMaggio, P. J. (Eds.). (1991). *The new institutionalism in organizational analysis*. University of Chicago Press.

Prinz, J. (2009). Is consciousness embodied? In: P. Robbins and M. Aydede (eds.), The Cambridge handbook of situated cognition. Cambridge: Cambridge University Press, pp. 419–37.

Prinz, W. (1990). A Common Coding Approach to Perception and Action. In: Neumann, O., Prinz, W. (eds) Relationships Between Perception and Action. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-75348-0_7

Przeworski, A. (2004). Do institutions matter? *In* J. L. Campbell, G. W. Carroll, J. R. Hollingsworth, & M. E. Streek (Eds.), *The emergence of economic sociology* (pp. 243-266). Princeton University Press.

Pylyshyn, Z. W. (1984). Computation and Cognition: Toward a Foundation for Cognitive Science. Cambridge: MIT Press.

Ramsey, W. M. (2007). Representation Reconsidered. Cambridge University Press.

Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2018). Answering Schrödinger's question: A free-energy formulation. Physics of life reviews, 24, 1–16. https://doi.org/10.1016/j.plrev.2017.09.001

Ramstead, M. J. D., Constant, A., Badcock, P. B., & Friston, K. J. (2019). Variational ecology and the physics of sentient systems. *Physics of Life Reviews*, *31*, 188–205. https://doi.org/10.1016/j.plrev.2018.12.002

Ramstead, M. J. D., Friston, K. J., & Hipólito, I. (2020). Is the free-energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy*, *22*(8). https://doi.org/10.3390/E22080889

Ramstead, M. J. D., Veissière, S. P. L., and Kirmayer, L. J. (2016). Cultural affordances: scaffolding local worlds through shared intentionality and regimes of attention. Front. Psychol. 7:1090. doi: 10.3389/fpsyg.2016.01090

Rietveld, E. & Denys, D., & van Westen, M. (2018). Ecological-Enactive Cognition as Engaging with a Field of Relevant Affordances: The Skilled Intentionality Framework (SIF).

Rietveld, E., & Kiverstein, J. (2014). A Rich Landscape of Affordances. *Ecological Psychology*, *26*(4), 325–352. https://doi.org/10.1080/10407413.2014.958035

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. Nature reviews. Neuroscience, 2(9), 661–670. https://doi.org/10.1038/35090060

Roberts, M. J. (2018). Signals and Systems: Analysis Using Transform Methods and MATLAB® (3 ed.). McGraw-Hill. pp. 131, 133–134. ISBN 978-0-07-802812-0.

Schelling, T. C. (1960). *The strategy of conflict.* Harvard Univer. Press.

Schieffelin, B. B., & Ochs, E. (Eds.). (1986). *Language socialization across cultures.* Cambridge University Press.

Schwitzgebel, E, "Introspection", The Stanford Encyclopedia of Philosophy (Fall 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2024/entries/introspection/>.

Scott, W. R. (2014). *Institutions and organizations: Ideas, interests, and identities* (4th ed.). SAGE Publications.

Searle, John R. (1983). Intentionality: An Essay in the Philosophy of Mind. Cambridge, UK: Cambridge University Press.

Searle, J. R. (2010). *Making the social world: The structure of human civilization.* Oxford University Press.

Searle, J. R. (2015). Status functions and institutional facts: reply to Hindriks and Guala. Journal of Institutional Economics, 11(3), 507–514. doi:10.1017/S1744137414000629

Sebanz, N., & Knoblich, G. (2009). Prediction in joint action: what, when, and where. Topics in cognitive science, 1(2), 353–367. https://doi.org/10.1111/j.1756-8765.2009.01024.x

Shannon, C. E. (1948), A Mathematical Theory of Communication. Bell System Technical Journal, 27: 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shapiro, L. (2011). *Embodied cognition*. New York: Routledge.

Smit, J. P., Buekens, F., & du Plessis, S. (2011). What is money? an alternative to searle's institutional facts. *Economics and Philosophy*, *27*(1), 1–22. https://doi.org/10.1017/S0266267110000441

Smit, J. P., Buekens, F., & du Plessis, S. (2013). Developing the incentivized action view of institutional reality. *Synthese*, *191*(8), 1813–1830. https://doi.org/10.1007/s11229-013-0370-5

Sperber, D. (1996). Explaining Culture: A Naturalistic Approach. Oxford: Basil Blackwell.

Sperber, D., & Hirschfeld, L. A. (2004). The cognitive foundations of cultural stability and diversity. *Trends in Cognitive Sciences*, *8*(1), 40–46. https://doi.org/10.1016/j.tics.2003.11.002

Sterling P. (2012). Allostasis: a model of predictive regulation. Physiology & behavior, 106(1), 5–15. https://doi.org/10.1016/j.physbeh.2011.06.004

Sterling, P., & Eyer, J. (1988). Allostasis: A new paradigm to explain arousal pathology. In S. Fisher & J. Reason (Eds.), *Handbook of life stress, cognition and health* (pp. 629–649). John Wiley & Sons.

Strogatz, S. H. (1994). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Addison-Wesley.

Sugden, R. (1993). Thinking as a Team: Towards an Explanation of Nonselfish Behavior. Social Philosophy and Policy, 10(1), 69–89. doi:10.1017/S0265052500004027

Teraji, S. (2017). Understanding coevolution of mind and society: institutions-as-rules and institutions-as-equilibria. *Mind and Society, 16*(1–2), 95–112. https://doi.org/10.1007/s11299-016-0196-1

Thyagarajan, T., & Kalpana, D. (2021). *Linear and non-linear system theory.* CRC Press.

Todorov, E. (2009). Parallels between sensory and motor information processing. In M. S. Gazzaniga (Ed.) The cognitive neurosciences. 4th edition (pp. 613–623). Cam- bridge, MA / London, UK: MIT Press.

Tomasello, M. (1999). *The cultural origins of human cognition.* Harvard University Press.

Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental Science*, *10*(1), 121–125. https://doi.org/10.1111/j.1467-7687.2007.00573.x

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*(5), 675–691. https://doi.org/10.1017/s0140525x05000129

Tomasello, M., & Rakoczy, H. (2007). What makes human cognition unique? From individual to shared to collective intentionality. *Intellectica*, *46-47*(2-3), 25–48.

Tschacher, Wolfgang & Dauwalder, Jean-Pierre. (2003). The Dynamical Systems Approach to Cognition.

Tuomela, R. (1995). *The importance of us: A philosophical study of collective intentionality*. Stanford University Press.

Tuomela, R. (2007). *Social ontology: Collective intentionality and group agents*. Oxford University Press.

Van Es, T., & Myin, E. (2020). Predictive Processing and Representation: How Less Can Be More. *The Philosophy and Science of Predictive Processing*, 7–24. https://doi.org/10.5040/9781350099784.ch-001

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience.* The MIT Press.

Veissière, S. P. L., Constant, A., Ramstead, M. J. D., Friston, K. J., & Kirmayer, L. J. (2019). Thinking Through Other Minds: A Variational Approach to Cognition and Culture. *Behavioral and Brain Sciences*, 1–97. https://doi.org/10.1017/S0140525X19001213

Vestner, T., Tipper, S. P., Hartley, T., Over, H., & Rueschemeyer, S.-A. (2019). Bound together: Social binding leads to faster processing, spatial distortion, and enhanced memory of interacting partners. Journal of Experimental Psychology: General, 148(7), 1251–68. https://doi.org/10.1037/xge0000545

von der Malsburg, C. (1981). The Correlation Theory of Brain Function. MPI Biophysical Chemistry Internal Report 81–2. Models of Neural Networks II (1994), Domany, E., van Hemmen, J.L., and Schulten, K. eds. (Springer). *MPI Biophysical Chemistry: Models of Neural Networks II, July 1981*, 1–26.

Vromen, J. (2003). Collective intentionality, evolutionary biology and social reality. Philosophical Explorations 6 (3):251-265.

Wiese, W., & Metzinger, T. (2017). Vanilla PP for Philosophers: A Primer on Predictive Processing. *Philosophy and Predictive Processing*, 1–18. https://doi.org/10.15502/9783958573024

Williams, D. (2018). Predictive processing and the representation wars. *Minds and Machines: Journal for Artificial Intelligence, Philosophy and Cognitive Science, 28*(1), 141–172. https://doi.org/10.1007/s11023-017-9441-6

Wittgenstein, L. (1953). *Philosophical investigations. Philosophische Untersuchungen.* Macmillan

Wu, Y. L., Lin, Z. J., Li, C. C., Lin, X., Shan, S. K., Guo, B., Zheng, M. H., Li, F., Yuan, L. Q., & Li, Z. H. (2023). Epigenetic regulation in metabolic diseases: mechanisms and advances in clinical study. Signal transduction and targeted therapy, 8(1), 98. https://doi.org/10.1038/s41392-023-01333-7

Zucker, L. G. (1977). The role of institutionalization in cultural persistence. In *American Sociological Review* (Vol. 42, pp. 726-743).