Preprints.org

Article

Not peer-reviewed version

Proto-Cognitive Bases of Agency

Fernando Rodriguez Vergara * and Phil Husbands

Posted Date: 17 October 2024

doi: 10.20944/preprints202410.1351.v1

Keywords: Agency; Proto-cognition; Causality; Autopoiesis; Autonomy; Enaction

Article

Proto-Cognitive Bases of Agency

Fernando Rodriguez Vergara * and Phil Husbands

AI Research Group, Department of Informatics, University of Sussex, Brighton BN1 9RH, UK

* Correspondence: f.rodriguez-vergara@sussex.ac.uk

† This paper is an extended version of our paper published in Artificial Life and Evolutionary Computation 18th Italian Workshop, WIVACE 2024, Namur, Belgium, September 11–13, 2024 (to appear)

Abstract: Autonomous systems are permanently in an ongoing interaction with the environment that surrounds them. Agency, in this sense, is often conceived as a system-environment asymmetry, where the system is able to influence the environment more than viceversa. Autonomous systems, however, are closed under information, meaning that they can only specify their own states by means of intelligent processes (unless we consider cognitive processes of higher order), even if they can influence and be influenced by physical processes outside of their mechanistic distinctions. Agency then, computationally speaking, should be understood as the causal effect that the system by means of its intelligent process can exert over itself, as in opposition to the environment. Being thus, we argue that a first measure of causality should capture the possibility of the system to withstand the effect of the environment over its future states, that is, to measure and compare the influence of the environment and the system, over the the system. We suggest that specific kind of mappings (represented as probability distributions) can increase the degree of underdetermination that can be quantified through information metrics (in particular, using entropy measure and the Earth Mover's Distance for comparisons) in the discrete case. After revising relevant related concepts, we introduce some mathematical formulations in the attempt to capture more formally the notion of underdetermination in autonomous systems. Finally, we exemplify these ideas and formulations through a toy-experiment in the the Game of Life cellular automaton

Keywords: agency; proto-cognition; causality; autopoiesis; autonomy; enaction

1. Introduction

Intuitively, agency can be portrayed as the capacity of a given system to determine, at least to some extent, its own behavioral trajectory (to 'decide' what to 'do'). Or, in other words, to display some degree of a causal decoupling from the world that both surrounds it and realizes it. Also intuitively, we often think of agency as an evolutionary trait or capacity, that would have arisen at a relatively late stage, after complex adaptive intelligent systems were in place. It seems somehow –at least from our experience, to be a natural solution to a type of problem for which more archaic forms of adaptive behavior (like pure structural adaptivity or reinforcement learning) are insufficient or less well fit; roughly speaking, any situation requiring some form of improvised responses and/or truly goal-directed actions, such as, for instance, 'simple' tasks like driving a car. This stage-like-leap is appealing because agency can be conceived as a higher degree of intelligence, stemming from more complex cognitive structures. Unfortunately though, beyond intuitions of this kind, our actual understanding of the underlying mechanisms supporting any real form of agency are still very unclear and have been a matter of increasing debate [1]. As [2] poses it, enactive cognitive science still needs to explain what would be the difference between an autonomous and any other kind of (intelligent adaptive) system. Hence, the existence of a system as a real autonomous entity such that it is embedded in an environment, but somehow causally is not just more environment itself.

Probably due to the historical philosophical focus around this topic (regarding thought), as well as to the initial anthropocentric scope of traditional cognitive science, agency has been mostly treated in terms of mental causation and/or the hypothetical neural capacities associated with it, therefore closely related with the notion of free will [3–8]. In the last decades, however, there has been a proliferation of theoretical approaches concerned with the logical principles of agency, that have switched the spotlight towards the fundamental properties enabling or endowing a system with it. These have taken a more abstract and functional perspective, inspired mostly by insights from theoretical biology and artificial systems research, permeating into several fields, such as, for instance, into philosophy [9,10], cognitive science [11,12], artificial life [1,13,14] and others [15,16], which has resulted not only in a reconceptualization of the notion of agency itself, but has also exposed a conceptual frailty of related relevant notions (e.g., what is a system in the first place?). Although pertaining to the same subject, given the novelty of the topic, plus the absence of a cohesive formal framework, it is possible to see quite dissimilar positions; for example, from agents as systems that act mechanically by consistently mapping its own states [17], to agency as a product of unpredictable neuro-physiological activity related to consciousness [18].

In this context, we propose to take a look back to the basics and start from the notion of selectivity. As we know, selectivity can be understood as a causal property that determines variable degrees of response specificity (depending on the complexity of the system, the environmental circumstances and the relation between them). Starting from this basic idea, we claim that, although physically co-specified by environmental circumstances, system state transitions in minimal autonomous (hence organizationally closed) systems can be characterized as dynamically more or less driven by an intrinsic 'steering force', underpinned by the degree of freedom that the organized sequential states can provide under the influence of different environmental circumstances. This is to say that, while the system's ongoing processes are unavoidably determined to an extent, a minimal form of agency will require a degree of under-determination in order to actively respond (as opposite to passively undergo viable transitions, hence depending only on an environmental "good-will"), this first condition arising from the structural mechanisms of the system will later enable enactions in a goal-directed fashion, therefore with an agential connotation. Simply put, while we know that both the states of the environment and the system will determine the upcoming state of the system, these contributions may not always be equal (less even an exact half-and-half), therefore naturally posing the question about the dynamics of this relation. Being thus, in the remainder of this manuscript we will explore this notion of a latent, proto-cognitive form of agency in terms of the asymmetrical causal contributions that the system and the environment exert (in spite of their inseparable nature) over the future states of the system and the possibility to formally approach a method for quantifying it.

2. Conceptual Bases

2.1. Fundamental Notions

In general, systems can be described in terms of their state dynamics, so that from any particular state and environmental conditions, a system will transition into a second state (which in some case may be some previously visited state, or even the same as the first). This is the basic, more intuitive, notion of the causal relation for how dynamical systems change.

In particular, autonomous systems belong to a specific kind of systems which are organizationally closed, meaning that their ongoing processes will recursively determine state transition from valid (i.e. viable) structural states into further valid states, and where the whole set of valid states that a system can undergo without disintegrating, will be what defines its (autonomous) organization [19–22]. As such, their actions (their state-transitions) are a consequence of their own organizational constraints, together with its environmental interactions and we can express these changes in terms of structural transitions of the kind: $(S_x, e_i) \rightarrow S_y$. Where S_x and S_y denote two consecutive states of the system, e_i the state of the environment and the $\rightarrow$ a causal relation or mapping between them.

It follows from this that, as long as the autonomous system remains so, neither the system, nor the environment can fully specify its next state (of the system). And, accordingly, we understand this process as a form of system-environment coupling that unfolds in time as the co-determined behavioral trajectories of autonomous systems. Autonomous systems, indeed, are fundamentally alien to such considerations; due to the unavoidable co-determination of their transitions, at least at basic level, they cannot *distinguish* system and environmental states as if they were-, or they belonged to- decoupled entities.

As a matter of fact, we know that a system can behave in the same way (i.e., perform the same state transition) even if the environmental conditions in which it is embedded are different, or, conversely, respond differently to identical external circumstances; these changes are not arbitrary and the underlying property we associate with this phenomena is the system's *selectivity*. The reason for this, is that selectivity depends on the specific structural characteristics of the system at the time of any given system-environment interaction, as this imposes limitations on the local interactions available to the *parts* of the system. Consequently, delimiting (i.e., determining a finite set of available state transitions underpinning-) its overall capacity for discrimination of external features and available responses. Therefore, any system will most likely portray a non-linear mapping among states, given its particular state-environment dispositions, especially in common cases of mismatching degrees of complexity, where injective mappings are impossible. Being thus, the variety of responses available to a system will be given, essentially, by the its structural states, inasmuch as these enable a higher variety of environmental distinctions and responses. Importantly, both of these are deployed as a single mechanism, which is what appears to an external observer as the behavior of the system [23]

More concretely, the fact that different states of the environment elicit identical system state transition entails that these environmental instances conform to categories, or sets of selectively equivalent environmental elements, such that:

$$\begin{aligned} (S_x, E_{xy}) &\longrightarrow S_y \\ E_{xy} = \{e_1, e_2, ..., e_n\} : (S_x, e_i) &\longrightarrow S_y \end{aligned} \tag{1}$$

Where every element e_i denotes some specific environmental state that belongs to the set E_{xy} whereby the same state transition $S_x \rightarrow S_y$ is enacted by the system. From now on, we will refer to these sets of causal and operationally equivalent system-environment interactions simply as equivalent categories.

While originally, according to the theory of autopoiesis [19,20] suggested that system-environmental interactions can be characterized by the notion of structural coupling, later developments on autonomy [22] and especially on autonomous cognition [23], led to criticism concerning the excessive rigidity of the initial autopoietic account and to its lack of a principled relation between components and the system they integrate as a whole [24–27]. This, in turn, led to enactions becoming the by-default notion when thinking about autonomous behavior, as well as one of the cornerstones of enactivism. As usual though, further research has brought more controversies into existence than the ones already present before and accordingly, there are several nuances regarding what is implied by the concept of enaction [28,29] and, of course, by agency [10,12,13,30]. To avoid confusions, by enaction, in our current minimal context, we shall understand an action/response that: *i*) is fundamentally a selectivity-based interpretation (under a plain mechanistic connotation) of the system-environment circumstances, thus, a distinction-action event mechanically performed by an autonomous system. That *ii*) is the primary component of the organization of the (organizationally closed) system while also organizationally constrained, and which is observable in the ongoing structural transformations of the system (its behavior). And *iii*) that it is causally self-referential, locally, as a causal consequence of the state system over itself (or rather its future state/self), whereas globally it can be abstracted as a causal consequence of the organization over itself, insofar as the organization is the cause and the result of the mapping among said states.

Coming back to Equation (1), we can more explicitly express an equivalent category as the set of environmental states $E_{xy} = \{e_1, e_2, ..., e_n\}$ by which the system transitions from a first state S_x into a subsequent state S_y, so that:

$$
\begin{gathered}
(S_x, e_1) \longrightarrow S_y \\
(S_x, e_2) \longrightarrow S_y \\
... \\
(S_x, e_n) \longrightarrow S_y
\end{gathered}
$$

Because state transitions can be expressed as dynamic mappings, which depend on the selectivity given by the particular state of a system (sometimes also characterized in terms of structural degeneracy), any relatively complex system will presumably exhibit a high variety of transition types. Some very simple cases to illustrate this, would be:

$$
\begin{gathered}
(S_u, e_i) \longrightarrow S_a \\
(S_v, e_i) \longrightarrow S_b \\
or: \\
(S_u, e_i) \longrightarrow S_y \\
(S_v, e_i) \longrightarrow S_y \\
or: \\
(S_x, e_i) \longrightarrow S_y \\
(S_y, e_i) \longrightarrow S_x
\end{gathered}
$$

For the first pair of transitions, the same environmental conditions will result in different states. In this case, the system's selectivity given by different structural instantiations is the cause of diverse responses. For the second pair, however, different initial states of the system will map onto the same state S_y under the same environmental conditions (a different state transition nonetheless). For this to happen, and assuming that $S_u \neq S_v$, the environmental state e_i has to be a part of both equivalent categories E_{uy} and E_{vy} even though the different states S_u and S_v entail $E_{uy} \neq E_{vy}$. Put another way, although their selectivity is different, their response is the same. The last two transitions characterize what would be a minimally recursive case, where in the absence of changes in the external conditions, the system will oscillate indefinitely between 2 states (but to exhibit a different selectivity as well). Indeed, this is a common case, for instance, in the Game of Life, where patterns such as blinkers or gliders oscillate between two configurations in the absence of perturbations. The contrast between the cases up to this point depict the conceptual role expressed by equivalent categories. Likewise, despite their simplicity, they are helpful to illustrate what we mean (minimally at least) by an enaction as an *interpretation* that is made by an autonomous system; a consistent categorization of system-environment coupled states through a coherent mapping of its own state transitions.

This implies that, on the one hand, all of the environmental conditions triggering, or co-specifying the same structural transition $(S_x, e_i) \rightarrow S_y$ are, from the point of view of the system, the same enaction (insofar as S_x maps into S_y and e_i is an element of the category E_{xy}). On the other hand, the behavioral complexity of the system (the causal range of actions available to a systems at any time and given some state S_x) is then the direct consequence of the causal structure of the set of all the equivalent categories available to it, which can be better captured as by a probability distribution. If, by abstracting away the effect of the environment, we consider the following open equivalent category:

$$
E_x = \{E_{xa}, E_{xb}, ..., E_{xz}\} \tag{2}
$$

Then, the sub-indices a, b, z will denote the possible future states of the system and E_{xa}, E_{xb}, E_{xz}, the corresponding equivalent categories and E_x the set of all the equivalent categories potentially available to the system in state S_x. And considering, that as long as the system can cope with external changes there might be as many environmental sets as the complexity of the organization of the system permits and, that the cardinality of these sets (given by the number of environmental states interpreted as equivalent) may differ greatly, then we can express the probability of the interpretation given by $E_x y$ and enacted by a transition into a state S_y, by the conditional probability distribution:

$$p(E_{xy}|E_x) = (p(E_{xa}), p(E_{xb}), ..., p(E_{xz})\} \tag{3}$$

At this point we are faced with a crossroad; either we take the probabilities in eq 3 to represent the chances of a state transition into S_y strictly on the bases of the environmental categorization, or we go further and investigate causality in terms of beliefs by examining the consistency of the interpretational mapping arising from them, as proposed by [17,31–34]. Although we believe the latter is a highly interesting approach, the present work is circumscribed within a very specific framework, namely, the proto-cognitive properties of autonomous systems. In this context, the kind of minimal directedness we are interested in is rather a primal form of aggregated selectivity, whereby an autonomous organization displays tendencies toward viable conditions (i.e. series of coherent structural transformations that make them interesting for us to analyze in the first place) that is necessarily strictly non-representational and non-phenomenological [20,35,36]. Thus, since the search for explanatory principles has to be cast, or so we suggest, in terms of a *blind* intelligence, or at least agnostic of biocognitive assumptions inasmuch as possible, here we will focus on filtering theoretical insights that involve explicit cognitive capacities.

Along these lines, what we are suggesting is to distinguish as proto-cognitive properties the specific intelligent properties that autonomous systems display and by autonomy a self-referential property [21,23,37]. Therefore a particular kind of intelligent behavior which is determined by organizationally closed dynamics and which is non-mental and non-organic. Put differently, a kind of intelligence that is logically and evolutionarily prior to life, while still constrained by a recursive nature, so that the coherences that a system exhibit are not just transient or evanescent processes, but safeguarded by being encoded in the structure of the entity that remains (a minimal autonomous system). In this respect, an enaction is proto-cognitive insofar as an intelligent action/response in a computational (albeit non-representational) sense. That is, a multiple realizable, intrinsically logical and consistent mapping of states that does not require for the system to be alive to be performed, matching other minimal accounts characterizing cognitive-like properties exhibited by artificial systems, paradigmatically exemplified through Bittorio in [23], but described profusely in the artificial life literature, see for instance studies by [38–41] to mention a few.

Along these lines, and as a final note before moving into the following section, we shall mention that whilst enactions actively modulate the dynamic selectivity of a (minimal) autonomous system, thereby determining its immediate disposition towards the environment, this is not the same –we will argue – as saying that they actively modulate its (relation with the) environment. Essentially, the latter operation entails more sophisticated cognitive capacities, probably involving some form of the aforementioned symbolic or phenomenological kinds of content [42,43] that we would prefer to avoid in the current setup.

2.2. *Agency and Autonomy*

Now then, a recurrent topic of theoretical debate has to do with the requirements for- and the mechanisms by which- autonomous systems can be said to become agents, or to exhibit traits of agency [1,10,13,44–46]. In general these are conceptually assumed to be posterior to the existence of autonomous systems as such and are often (even if several nuances are in place) conceived as arising from some asymmetric system-environment relation, whereby the system somehow is capable of influencing its environment to a higher degree than from the *opposite* causal direction (environment over system) [9,11,25,30,47].On the one hand, the very notion of agency entails that some primal form of asymmetry is necessary to functionally decouple a system from its environment. On the other hand, we believe that the notion of a system-environment modulation, at least in the context of non-biological minimal autonomous systems –hence in proto-cognitive terms, is not actually possible. In this section we will unpack the reasons for this.

Let's start by, in spite of its simplicity, looking at Equation (1), from where we can see that enactions express, fundamentally, as system-environment co-specifications of the subsequent states of the system. Clearly, insofar as any mutual cause-and-effect event, there will be a counter-effect produced by the existence of the system that plays the role of the environment's environment, however, this influence of the system over the environment is not different in principle than that of any two non-autonomous systems interacting, like wind eroding a rock (assuming a non organizationally closed environment, if not the relation is one of bidirectional/social adaptation). Of course this could be just the formulation being wrong. However, what we'd like to point out is that the physical bidirectional unfolding by which system and environment influence both their future states reciprocally, must be something different in nature for systems that are organizationally closed than for those who are not; otherwise, any form of agency would be precluded on physical bases from the start.

In relation with this last point, note that agency in our context could be perhaps linked to another matter of debate, namely, the relation between agency and consciousness, however, the scope of the present work is a different one. Basically, these accounts often posit this relation in two ways: a consciousness-to-agency direction, that builds up from the fact that in spite of our empirical experience, our intuition about agency (i.e., as conscious decision-action actions) seems to clash against our understanding of the physical laws of the world, to propose that agency might somehow spawn from experience [18]. Or in an agency/causality-to-agency direction, by which causal relations would be the source of phenomenological experience [48] (see, for example, [49–52] for some interesting discussions). Although relevant, we deem these approaches to be problematic for an account of agency, centrally because of their grounding in the human case, by which there is an unavoidable preconception, thereby not only introducing the problem of formally characterizing phenomenological properties (an even harder enterprise, considering the current lack of philosophical and scientific understanding in this respect), but also because agency from a higher-level human stance involves too many confounding psychological and social factors that obscure the possibility of a minimal characterization.

On a similar vein, and this is the crux of the present argument, most enactive accounts of agency as such (i.e., regarding its fundamental principles and hypothetical mechanisms), have implicitly portrayed it as a function of higher order cognitive traits, hence associated with living beings, generally construed as (more or less) intentional decision-making capacities underpinned by adaptive responses arising from their precarious nature [25,30,44,53]. Although this may be ultimately right, it nonetheless forces agency to be portrayed as a mental construct, by framing it in terms of other theoretical notions that demand further assumptions on symbolic or phenomenological capacities required to *make-sense* of the environment [42,43], hence, discarding beforehand the possibility of minimal forms of agency that could be present in non-living autonomous systems.

To better understand this, we shall provide a bit of context about two important issues; the first pertains to the passage from the initial account of environmental interactions of autopoietic systems (i.e., structural coupling) [19,54,55] to the later notion of enactions. The second, concerns the related notion of structural determinism posited by [20] which unfortunately lost relevance with time.

While in its early conception of autonomy [21], autonomous organizations were already defined as a subset of the organizationally closed ones, the reasoning behind this was still rather abstract. A common example is the set of all natural numbers, which can be thought as an organizationally closed domain under addition and multiplication, insofar as any operation over any element will produce another element that belongs to the same domain. This entity (the set of natural numbers) however, is not autonomous; it doesn't *act* on its own because its organization is immutable and it cannot be perturbed. This is why every autonomous system is organizationally closed, but not viceversa. Crucially, the existence of an environment will therefore imply the step from organizational closure to autonomous dynamics.

This idea reshaped the notion of autonomy into its later enactive form [23], which heavily relies upon a system-environment coherence and where cognition is explored in the light of an embodied and embedded view of autopoiesis, autonomy and centrally through the concept of enaction. In this later (enactive) view, the logic underlying autonomy arises as a consequence from a required system-environment coherence, where system and environment continuously co-determine the behavioral trajectory of the system. This shifts the view of the system from a *passively* drifting machine into an active (cognitive) participant, implied by the notion of enaction as a (perceptual) distinction and action as a single process or cognitive event [29]. Otherwise, in spite of its internal coherence, the system would eventually (and too easily) disintegrate under the influence of a flux of unpredictable external changes [23–25].

Enactive principles (at least until this point) are still mostly aligned with previous ones from autopoiesis [19] and may even be understood as essentially equivalent regarding how the autonomous logic of a system dynamically determines the instantiation of a structure with a certain selectivity that is organizationally constrained by the conservation of the organization itself, meaning that it can engage in some interactions (but not others) in some domain; with the exception, of course, of the purported active role of the system. Nevertheless, the emphasis on the continuous system-environment interaction eventually leads to a sort of reification of environmental interactions, as a if the environment were a cognitive object (from a hypothetical 'point of view' of the autonomous system). Which would the imply that, somehow, the system is capable to *conceive* and monitor the conditions for its own persistence [43], or to *conceive* and manipulate its environment in an intentional (mental or proto-mental) fashion [42], or both.

The conceptual distinction here is slight but important, because structural/behavioral adaptation, for instance through negative feedback loops, does not entail actively modulating the environment with some explicit goal, not even if this goal seems to be the conservation of itself (as seen by us as external observers). This requires a form of intelligence far beyond basic structural adaptive capacities (plasticity possibly included, but fundamentally without complex nested or massive interdependencies), by which the system could discriminate what itself is (its boundaries, its processes, or something along these lines) and what is not itself (i.e., its external environment in the simplest case), which is not evident from what we know of the dynamical properties of minimal autonomous organizations (this becomes more clear if we think of Bittorio [23] or the patterns in the Game of Life [56,57] as basic examples). It's on these grounds that we claim that positing and characterizing agency as an asymmetrical system-environment causal relation, at least at a minimal level, is conceptually flawed.

The relevance of the second issue we anticipated, that is, the notion of structural determinism, is that it offers a conceptual alternative to the aforementioned problem. In the last part of this section we shall briefly introduce this idea, along with explaining its potential usefulness.

An important property derived from the organizationally closed nature of autonomous systems is their closure to information. As noted by [58], these systems are of the same kind as those described by [59]: thermodynamically open (insofar a their ongoing processes are based on a constant flux of energy exchange with their environments), while closed to information, given that they specify their states in a self-referential fashion (their behavior is not –it could not be – externally specified from the environment) [19–21,55,60]. This is the principle underpinning the idea that autonomous systems have no inputs or outputs [58,61] and the reason by which the more classic notion of determinism is replaced by one of structural determinism, whereby the causal relations guiding the operation of the system are conceived as intrinsic, inasmuch as the causal effects of the environment over the system, while certainly not irrelevant, are primarily a consequence of the system's structure [19,20,42,58]. In simple words, the system is able to specify its own behavior, because is equally unable to do otherwise; each structural change will be determined by its own (organizationally determined) structure. The environment exists, of course, but it is inaccessible as a source of objective information.

Information, in this context, may be better conceptualized in the sense of instructions. Indeed, because any and all instructions regulating the behavior of an autopoietic system are intrinsic to its endogenous processes, there cannot be instructions from the outside. This is why different system respond differently to the same external conditions, because autonomous systems cannot receive objective specifications (instructions) directly from outside, even if they wanted to. Conceptually speaking, system-environment interactions do not specify the behavior of the system, insofar as it is the structure of the system what defines what the environment is to it, thence its subsequent responses [42,54,58]. Furthermore, because the organization is closed, any interpretation a system can make from its interactions will be subjective (i.e. mechanistic, albeit particular to it) and relative to the nature of the (autonomy) of the system.

As a matter of fact, given their self-referential dynamics, organizations of this kind may reshape the topology of their organizational state-space (by forming attractors or *discovering* new state-transitions, but not the underlying intrinsic logic (because their logic *is* the identity of the organization itself). To continue with our previous example; if we were to somehow embed the set of all integer numbers in an environment in a way that motivated spontaneous operations, its organizational closure would hold as long as the logic determining transitions (i.e., the operations: addition, subtraction and multiplication) remained so. (This is why organizational and operational closure are frequently used as exchangeable terms, because the organizational state-space is closed under specific operations). The same applies to the validity of the system's states, which would correspond to the nature of the numbers themselves (e.g. excluding imaginary instances, as this would disrupt the overall dynamical coherence even under the same logical operations). The point that we'd like to emphasize is that, even if some external force can elicit changes, these changes are *directed* not towards an hypothetical representation of the environment, but to its own future states, because as long as the organization holds, their causal influence will be confined to the state-space of the organizational dynamics. Systems of this kind are self-referential in an absolute sense, therefore they are devoid of any notion of the environment (or of themselves, for that matter), hence precluded from valence-endowed impressions that could guide their responses. In this sense, the traditional denial of (cognitively objective) information in cognitive science [19,29,44,45,62] does not really clash with the more established notion of information nowadays, by which we basically understand a measure of correlation between observable phenomena.

As stated above –now with enough theoretical context in place, the goal of this manuscript is to propose that a fundamental element for the operational substrate for what we conceive as agency is already present, as a latent proto-cognitive property of autonomous systems, underpinned by their recursive nature and the degree to which the influence of the system over its own states can withstand that of the environment. This, we suggest, is the primal form of asymmetry required for agency (in its further cognitive connotation) and it depends on the extent to which such future states (of the system) are a causal consequence of the effect of its own organized unfolding. In this sense, we would like to pose a dynamic view on the system-environment coupling, in which the effect of the environment over the system is not always straightforwardly given, but it may vary depending on whether the system is able to, at least occasionally, specify states that could produce divergent subsequent states (so, alternatives) under the same environmental constraints, closer to the notion of degrees of freedom.

We do not intend to claim that the system could cause (i.e., fully determine) its states to come, but that there is a *weight* than can be measured, over the assumption that both contributions (system and environment) interact to determine system behavior. Hence, on the basis that constitutive and interactional properties relate to each other [26,27]. In fact, in spite of specifying itself, the process of self-determination is not context free and not univocal either. Moreover, in terms of raw causal power the asymmetry is biased on the other direction; it is only the environment that can on its own determine the future of the system (like its death from conditions the system cannot counter, considering an extreme case) because only the autonomous organization has its organizational closure to preserve. Put another way, the structure of the system will respond differently to different environments depending on its own structure, but whatever response it produces, it will also occur in a certain environmental context that logically precedes (is necessary for-) its autonomy. Hence, at this level, the causal *power* of the organization can be understood as an informational (instruction-wise) membrane allowing (1) a certain degree of environment-to-system under-determination and, (2) a certain degree of *specious* redundancy through equivalent categories, enabling multiple (or at least more than one) potential responses to the same environmental conditions.

3. Causality as Information

3.1. Intrinsic Information

Building on Bateson's idea of "differences that make a difference" [63] and along similar conceptual lines about selectivity that we have examined until now, the Integrated Information Theory of Consciousness (IIT) [48,64–66] introduces the formal notion of intrinsic information, in order to quantify the causal effect a system, given its structure, has over itself. Roughly speaking, the IIT claims that integrated information, a measure of emergent causality (often also expressed in terms of causal power), reflects the degree to which a given system intrinsically exists as such (so independent of external observation), hence, as a causally decoupled entity that becomes a phenomenological observer itself [67,68]. Technically, the goal of the IIT is to provide a formal framework for quantifying consciousness from physical mechanisms, under the premise that the ontological existence of the system entails a phenomenological identity [48,67], under the broader paradigm of causal emergence [69].

Basically, intrinsic information is the idea that the causal links of an irreducible system have to be self-contained. In turn, cause-effect information is the purported measure for the degree of causal-power exerted, *within* and over itself, by the very existence of a system/mechanism (and a hint of the implied intrinsic subjective observer), given the state transition probabilities from past and to future states, applied to the minimal mechanisms (units) of a candidate system [48]. On these grounds, causality becomes fundamentally linked to selectivity, which is understood as informative (with a connotation of specificity), insofar as the state of a system and its mechanisms logically specify a finite set of possible past and future states while discarding others [48,66,70,71].

From there, integrated (intrinsic) information (ϕ) is a second step; an attempt to formally solve the combination problem [72] through the notion of an causally emergent unity, which would be *locus* of phenomenological experience of the whole system, assuming that each of the elements of the system could have some degree of phenomenological experience on its own, but that they *integrate* into an irreducible physical/phenomenological mechanism. Simply put, integrated information is be the aggregation of what intrinsic information reveals.

Although the theory has gathered a lot of interest, it has also received plenty of criticism; pointing to its lack of a principled justification for the leap from maximally integrated information to consciousness [73–75]; inconsistencies among different measures for integrated information [71]; unconvincing mechanism accounting for temporality, especially regarding the exclusion principle [76–78]; and the apparent impossibility of consistently apply its methods to more complex systems [79,80], among others. In this respect, while we don't commit to the axiomatic postulates of the IIT [48,66,67], nor believe there is a strong enough reason to accept the proposed causality-consciousness identity, we do praise the attempt of orienting the discussion into a scientific domain. Moreover, we believe that, as others have done by building on some of these ideas to approach particular problems, usually with respect to the linked notion of emergence [81,82], the application of the specific notion of intrinsic information may be quite fruitful in the context of our present investigation about agency, under a strict selectivity-causality interpretation. We will henceforth develop this idea.

To this end, we will make use of the cause and effect repertoires introduced in [66]. IIT repertoires assign a markovian probability to every possible state transition, thereby producing two probability distributions; a cause (past-to-present) repertoire (ci) and an effect (present-to-future) repertoire ei.

$$ci = EMD\left(p\left(S^p|S^c\right) || \, p^{uc}(S^p)\right) \tag{4}$$

$$ei = EMD\left(p\left(S^f|S^c\right) || \, p^{uc}(S^f)\right) \tag{5}$$

Where the terms S^p (past), S^c (current) and S^f (future) refer to the system states, considered in the context of subsequent timesteps (hence, equivalent to S^{t-1}, S^t and S^{t+1}). *ci* and *ei* stand for cause and effect information respectively, and *cei* for cause-effect information quantifying the causal power measurable from the system. EMD stands for the Earth Mover's Distance [83,84] which is applied to compare cause/effect repertoires against non-intrinsically causal, unconstrained conditions (denoted by the superscript uc) to provides a measure of how much the system determines its own behavior (state transitions) away from random changes or pure external determination, such as ordinary entropy decay. From these, cause-effect (intrinsic) information is obtained by taking the minimum between them: $cei = min(ci, ei)$. The minimum reflects the shared degree of causality that we can ascribe to its structure in both directions.

Roughly speaking, the notion of cause and effect repertoires can be simplified and illustrated as in Figure 1, where the sets S^{t-1} and S^{t+1}, both contain all the possible states of the system, so that:

$$S^{t-1} = S^{t+1}$$

$$\{x_1, ..., x_n\} = \{z_1, ..., z_n\}$$

And where the multiple arrows connect each of these states to the current state of the system (represented by y). Hence, in this case, IIT repertoires assign a probability to each of these arrows, thereby producing two weighted mappings; a cause repertoire, which assigns a probability related to every possible (past) state of the system at $t-1$, and an effect repertoire which does the same for the upcoming (future) states at $t+1$.

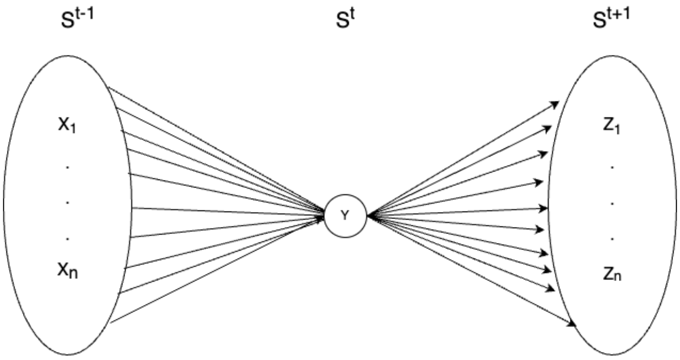


Figure 1. A simplified illustration of cause and effect repertoires. The sets S^{t-1}, S^t, S^{t+1} represent past, current and future states respectively; repertoires assign probabilities to each of the arrows. See main text for further details.

3.2. Cause-Effect Information in the Game of Life

Moving onto our goal at hand, we will first test whether cause-effect information can be applied to approach a measure of causality (in the sense of agency we have discussed) by applying it to the simplest case to be found in the Game of Life; namely, a single cell with only two states and its environment, made of the surrounding 8 cells. This space, containing a central cell plus those at a Chebyshev distance equal to 1, is also known as the Moore neighborhood [85] and in this particular case can display $2^9 = 512$ different configurations ($2^8 = 256$ for the environment). From the rules of the GoL [57,86,87] we know that a cell can only be 'alive' (active) if it is already active and there are 2 or 3 active cells in its Moore neighborhood (apart from the central self itself), otherwise, if not currently active, only if the sum of active cells in its surroundings is exactly 3. This can also be expressed as:

$$C_y = \begin{cases} 1 & \text{if } C_x = 1 \text{ and } \sum N(C_x) = 2 \vee \sum N(C_x) = 3 \\ 1 & \text{if } C_x = 0 \text{ and } \sum N(C_x) = 3 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Where $N(C_x)$ stands for the neighborhood (the Moore neighborhood minus the central cell) for a central cell (C_x) that follows a state transition $C_x \to C_y$. Note that the neighborhood in this case, corresponds to $E_x(C_x)$ (the equivalent categories available to C_x) instead of e_x (the actual environmental configurations), because for the central cell, the specific states do not make any difference, only their sum. And since there are: 28 combinations for $E_x = 2$, 56 combinations for $E_x = 3$ and 172 combinations for the remaining non viable (i.e. deactivation) alternatives, we then have the following counts:

$$C_x \to C_y = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \left\| \begin{array}{cc} 200 & 56 \\ 172 & 84 \end{array} \right\| \end{array} \tag{7}$$

From where we can derive the probability matrices:

$$T(C_x) = \begin{vmatrix} 0.78 & 0.22 \\ 0.67 & 0.33 \end{vmatrix} \quad ; \quad T(C_y) = \begin{vmatrix} 0.54 & 0.46 \\ 0.4 & 0.6 \end{vmatrix}$$

Now then, we can follow the IIT formulation (from equations 4-5) and calculate information, by first computing cause and effect repertoires based on the current state of the cell. For the sake of the example, we will start with an active central cell at a given time t, $C_x = 1$:

$$crep(C_x = 1) = p(C_x^p \mid C_x = 1) = (\frac{56}{140}, \frac{84}{140}) = (0.4, 0.6)$$

$$erep(C_x = 1) = p(C_x^f \mid C_x = 1) = (\frac{172}{256}, \frac{84}{256}) = (0.67, 0.33)$$

Where *crep* and *erep* stand for cause and effect repertoires and the super-indices ^{p} and ^{f} refer to past and future states respectively.

We can also obtain the unconstrained (past) UC^p and (future) probability distributions UC^f through direct counting. First, by considering that UC^p represents the probabilities of the past state of the system without any knowledge of its current state (transition into C_y if looking at the transition counts in (7)), so unconstrained by it, which logically entail a uniform distribution.

Similarly, UC^f represents the probabilities of the future state of the system without any causal input from (or, again, unconstrained by) S_x, which is the same as not having any knowledge about the current state of the system, hence a vertical sum of the elements in the transition matrix from Equation (7).

Put differently, whereas the unconstrained past is formalized as a simple homogeneous distribution (assuming unconstrained outputs), the unconstrained future distribution is taken as independent of the current state of the system, although still dependent on its causal structure (i.e. as a system with unconstrained inputs). In this sense, the unconstrained repertoires represent the marginal probability mass in both directions. Formally, this can be concretely expressed as:

$$UC^p = p(\sum_{i=1}^{n}(C_{y_i}|C_x = 0), \sum_{i=1}^{n}(C_{y_i}|C_x = 1)) = (0.5, 0.5)$$

$$UC^f = p(\sum_{i=1}^{n}(C_{x_i}|C_y = 0), \sum_{i=1}^{n}(C_{x_i}|C_y = 1)) = (0.73, 0.27)$$

Moving on, cause and effect information are computed by comparing cause and effect repertoires against their unconstrained (UC) reciprocal repertoires. Information is calculated in terms of distance (difference) between the compared distributions, using the EMD:

$$ci = EMD(crep(cell = 1) \mid\mid UC^p) = EMD((0.4, 0.6) \mid\mid (0.5, 0.5)) = 0.1$$

doi:10.20944/preprints202410.1351.v1

$$ei = EMD(erep(cell = 1) \mid\mid UC^f) = EMD((0.67, 0.33) \mid\mid (0.73, 0.27)) = 0.0547$$

Thus $cei = min(ci, ei) = 0.0547$. Which, leaving aside hypothetical connections to phenomenological properties of any kind, is basically an expression of the fact that our knowledge about the ON state of a single cell is informative to the extent it give us an insight into its past and future states. From this it follows that, when put in terms of causality, cause-effect information may be interpreted as the constraints that the state of the ON cell places upon its state transitions, hence the degree of self-determination with respect to its environment (i.e. the rest of the Moore neighborhood, within the context of our example).

Then, by repeating the process for $C_x = 0$ we obtain:

$$ci = EMD((0.54, 0.46) \mid\mid (0.5, 0.5)) = 0.0376$$

$$ei = EMD((0.78, 0.22) \mid\mid (0.73, 0.27)) = 0.0547$$

Then making $cei = min(ci, ei) = 0.0376$. This, as might be expected, shows that active cells on the grid have a slight higher causal power (0.017) than non-active cells, which is reasonable considering the amount of transitions leading to ON and OFF states by the dynamics of the Game of Life.

As a recapitulation from our toy case example we can highlight some simple things; first, information is low, because from only knowing the value of one cell we don't really have a lot of information about its past or future state, as this depends mostly on its neighborhood. Second, the highest and lowest information values correspond to the cause information for an active cell ($ci(C_x = 1)$) and ($ci(C_x = 0)$) respectively, meaning that in the context of the GoL, single active cells are the most informative and, conversely, non-active cells provide the least information when looking backwards, which fits in with what we know about its dynamics as well. Lastly, do note that the fact that the value of effect information is the same in both cases is not due to identical effect repertoires, but a result of the geometrical equivalent distance with respect to the unconstrained probability distribution for future states. Although these simple points are not really significant, they are helpful to emphasize the possibility of causal interpretation from the application of the IIT formulations, without further conceptual escalation into more uncertain matters.

As a final note in this respect, it may be important to mention that there is no need for further integration calculations or analysis in that regard, because the object of study we have selected (a single cell on the grid) is, by definition and by the dynamics of the GoL, the minimal possible (therefore irreducible) case.

The point that we will like to make visible is that, in spite of the potential problems with the purported conceptual implications we have expressed regarding the IIT (which we will briefly discuss again in the following section), insofar as measures of an intrinsic property of the system in terms of information, not only they are quite insightful, but also well fit for further application (for a good review and example of this, in the context of emergence, see [82]). Being thus, in the remaining of this chapter we will take this approach and elaborate over the original formulations, to derive a different formal construct that, so we propose, can be applied to measure protocognitive underdetermination.

4. Under-Determination as Latent Agency

4.1. Some Considerations

Aside from the general criticisms to the IIT that we mentioned at the beginning of the previous section, there are a few specific aspects that, we believe, require further examination. Specifically, regarding the notion of intrinsic information; we believe that there is the need for a conceptual switch from a single state-based minimal unit, to an enaction (in the very minimal sense used in Varela et al. [23]). We shall briefly discuss this now.

As it has been mentioned, the *subjective* (strictly mechanically speaking) interpretational dimension of autonomous systems stems from two intertwined features; their dynamic structural selectivity and their (organizationally) coherent state transitions. This is what underpins their (at least proto-cognitive) intelligence and, therefore, propels their consistent distinctions-actions. Neither selectivity, nor intelligence, have a mental or even cognitive connotation in this context of course, but they serve as mechanistic concepts to understand the dynamic evolution of a particular kind of system, so much so that, often, even biological systems disintegrate or dissipate on the bases of their selectivity.

Nonetheless, the reason that cognitive theories (or other approaches to cognitive phenomena, such as the IIT) stress the importance of this notion, is because any reliable cognitive distinction made by a system will require a stable pattern of structural change that correlates with specific traits of the environment, as otherwise, in the absence of this minimal form of consistency, there would be no cognition and no observer whatsoever. Furthermore, structural selectivity can only be conceived as an intelligent/cognitive property inasmuch as the system upon which it's operating is operationally closed. Since any structural change that could be considered to be a distinction, an action, or anything along those lines, could only be so if, and only if, there is some entity in the first place. In other words, there must be something (some organized system) that will structurally change, while at the same time remain being the same (organizationally speaking), to which the *difference that make a difference*, or any difference whatsoever (for what it counts) entailing an intelligent distinction instead of another has any relevance at all (again, strictly mechanically speaking).

It follows from this, or at least so we argue, that in order to correctly account for what a system does, we need to understand what a system interprets and vice-versa, in terms of minimal distinction-action (proto)-cognitive responses, therefore in terms of enactions. Enactions are, in effect, proto-cognitive irreducible events and, as such, they bring a temporal component that is absent from markovian analyses; rather than possible transitions being informative in relation to a state, the transition is in itself informative as a whole insofar as the change it creates, which discards the possibility of other changes (i.e., perceptually guided behaviors). Summarizing what we have argued up to this point, we can say that:

i) State transitions in autonomous organizations are not only causally related to action, but to both; perception and action as a single process that is a material manifestation of some system-environment coherence (i.e. enactions). Enactions involve (in the minimal discrete case) a pair of system states, but also an environmental interaction that *impels* certain dynamics, and a consistent mapping function between them, which is encoded in the structure of the system itself. This is a process, instead of a state, therefore extended –even if minimally, in time.

ii) Information-wise, autonomous systems are closed (self-referential), where information has the connotation of instruction or specification. It is necessary to distinguish then, causality in the matter/energy sense in which system-environment interactions are open, and in the former sense, in which even individual enactions may be considered informationally closed.

iii) Although agency is intuitively understood as pertaining to the energy/matter dimension, given that organizationally closed systems can only computationally (i.e., by means of intelligent processes) specify their own states, whatever else they causally influence or are influenced by-, by means of their material states and changes may be certainly causal, but it is not a proto-cognitive feature of the system, because information is confined to the organization itself.

iv) The consistent mapping between the first system-environment and the second state is the most basic form of coherence displayed (continuously) by an autonomous organization. Our contention is that the internal structure of these mappings can be more complex than a simple injective function and, inasmuch as this mappings are not univocal, they represent an interpretational space, which is better expressed through probability distributions.

v) Agency, then, is not a property that somehow could subvert the matter/energy order, but rather –at least at the level of autonomy – is one that gives rise to and amplifies the degree of under-determination underlying the operation of the system. We termed this particular quality *latent* agency, to stress the fact that the exploitation of this increase of under-determination will require higher order, representational or phenomenological properties to render such actions/responses *meaningful* to the system/agent.

In concrete, we will move away from the past-current-future view by examining only the two states of the system involved in any system transition. By recalling the formulation of repertoires from Equations (4) and (5), we can express a single causal direction as:

$$rep_{xy}(S_x) = p(S_y \mid S_x = X) \tag{8}$$

Here, rep_{xy} corresponds to the effect repertoire, which, while conceptually different, is formally the same (X represent the specific state of S_x). The notation has been modified to avoid unnecessary confusions further on.

In order to investigate possible asymmetric contributions, we need to conceptually delimit system-system and environment-system influences and try to formally disentangle them, for quantification. As we have discussed system self-determination effects until now, we will focus on the elaboration of the environment-system side of the coin.

Tentative starting points could have been the enactivist notion of intentional (in the sense of directedness) action [88,89] or interactional asymmetry [11,44]. However, as we discussed above, we'd like to avoid cognitive properties of higher order than minimal autonomous systems. Regarding the latter, while we ascribe to the proposed connection among self-individuation, normativity and asymmetry as requirements for agency, especially considering that –at least from our perspective, all of these properties can be theoretically described in proto-cognitive properties stemming from autonomy; we also wish to avoid the theoretical system-to-environment influence described as an asymmetrical modulation.

Essentially, this would depict a notion of agency that seems somehow unfeasible, either without representational capacities (e.g., requiring a model of the world) or some minimal phenomenological attributions (by which some form of *meaningful* sense-making could drive the system modulation of the environment in accordance to its needs) [36,43,78], which would implicitly entail that even minimal manifestations of agency are logically dependent on biological or most probably higher cognitive correlated properties.

Being thus, we will leave aside from our constructs the idea of the future-environment element and incorporate the notion of asymmetry only in terms of the influence that the environment has over the system, in order to enable a formal method to later examine it in opposition to that of the system upon itself. To this end, we will introduce a second repertoire:

$$rep_{ey}(E_{xy}) = p(S_y \mid E_x(S_x = X)) \tag{9}$$

Where the term E_x refers to the equivalent categories available to the system, S_x to the state of the same system and X to some specific state. The argument is E_{xy} because rep_{ey} is a weighted mapping of the equivalent category defined by S_x and S_y. Thus, it expresses the probability of the next state of the system, considering the interpretations that the system can make from the conditions of the environment. In other words, this is the effect of the environment over the system, given its particular structural selectivity; the proto-cognitive interpretation of the environment that the system makes.

Having both repertoires, we are now provided with the elements to examine causality by quantifying it in terms of information. Along these lines, the first step will be to compute the entropy of these individual repertoires, where entropy is defined as:

$$H(X) = -\sum_{x \in X} p(x) \log_b p(x)$$

Entropy, in this context, can be understood as an indicator of causality, or maybe better, an indicator of the latent degree of causality allowed.

Later, along the lines of intrinsic information, we will make use of the EMD metric to compute the distance of the repertoires against ideal conditions (highest and lowest entropy cases respectively) and then between them. In this case, given that the Earth Mover's Distance is a commutative metric, distance would reflect how far they are from each other, although without indicating which of the repertoires is causally stronger than the other. Being thus, we will apply EMD measurements three times: one to compare the environment-system influence (rep_{ey}) against a uniform distribution (which in theory would be the one more exploitable by a system). Another to compare the system-system repertoire (rep_{xy}) against a totally determined distribution, and a last one to measure the distance between the repertoires themselves, hence hinting at how stronger the causality is from one with respect to the other, which along with the rest of the measurements will give us a relatively good idea of the general picture. We will develop these ideas in detail in the next section.

4.2. Coming back to GoL

Coming back to the minimal case of a single cell in the Game of Life and following from the descriptions given in the previous section, we can note that, whereas the objective environmental states conform a set encompassing all the possible configurations that the 8 cells in a neighborhood of the central cell can instantiate, the equivalent sets that the central cell can distinguish are only 3 and given by their sums. Hence, we can group these cases and get the number of combinations by applying a simple combination formula:

$$C(n,r) = \binom{n}{r} = \frac{n!}{(r!(n-r)!}$$

From where we obtain: $C(8,2) = 28$, $C(8,3) = 56$ for transitions into active cells (so that $C_y = 1$) and $C(8,q) = 172$ for the remaining cases. The symbol q refers to the rest of the sums: $E_{xy} \neq 2 \vee 3$, whereby only transitions into non-active states are generated. From this, we can systematize transitions in terms of enactions as in Table 1.

Table 1. Table displaying the possible system-environment combinations and their resulting state transitions, for a single cell in the GoL.

(S_x, E_{xy})	$C_y = 0$	$C_y = 1$	Total
$(0,2)$	28	0	28
$(0,3)$	0	56	56
$(0,q)$	172	0	172
$(1,2)$	0	28	28
$(1,3)$	0	56	56
$(1,q)$	172	0	172
Total	372	140	512

We have previously examined this system on its own, hence, we now need to derive the possible transitions as a function of the environment (i.e., the environment-system component of the coupling). This is presented in Table 2.

Table 2. Table displaying only environmental states encountered before transitioning and their correspondent resulting states. Hence, as environment-to-system influence

E_{xy}	$C_y = 0$	$C_y = 1$	Total
2	28	28	56
3	0	112	112
q	344	0	344
Total	372	140	512

After this, by making use of Table 2, we finally are in position to derive the environmental causal components as probability distributions (repertoires). As it can be inferred, the only environmental category that enables some degree of causal freedom to the system is $E_{xy} = 2$, because for all cases in which the sum of the surroundings cells is 3, the subsequent state of the single cell in the centre will unavoidably be ON ($C_y = 1$), independently of its own state. Conversely, for any other case, the subsequent state of the single cell at hand will be OFF ($C_y = 0$). This exemplifies different degrees of causal environment-state determination and can be related to the entropy of these distributions (in bits):

$$H(rep_{ey}(E_{xy} = 2)) = H(0.5, 0.5) = 1$$

$$H(rep_{ey}(E_{xy} = 3)) = H((rep_{ey}(E_{xy} = q)) = 0$$

Therefore, illustrating the two extreme cases; the former shows how for some environmental categories, the selectivity of the system allows a degree of under-determination that may be exploited by the system. Conversely, the other two cases totally preclude any kind of under-determination. In both cases the system is self-determined, the key difference is that, for E_{xy} what determines the state transition (so the enaction) is the state of the system (C_x). The entropy of the correspondent distributions given by rep_{xy} can be calculated accordingly:

$$H(rep_{xy}(C_x = 0)) = H(0.78, 0.22) = 0.76$$

$$H(rep_{xy}(C_x = 1)) = H(0.67, 0.33) = 0.914$$

Entropy, as an indicator of the uncertainty of the outcome given by the probability distribution (as it can be seen from the extreme cases of the entropies resulting from the environmental repertoires) has two important interpretations: Regarding the environment, as we have seen, entropy tell us the degree to which a system might be able to exert causal effects in its own future states, if any at all. This seems intuitively easier to understand it the other way around (system-to-system); in this case, a probability distribution with very uncertain outcomes (like one close to a uniform distribution), would probably provide a smaller possibility for a (higher order) system to influence its future states. Because, for this to occur, it will need not only to generate environmental under-determination, but also to exploit it, and the more focused tendencies will be result in lower entropy values. Therefore, likely ideal conditions would be, at least generally speaking, high environmental interpretability and entropy, along with focused, so low, system entropy. Simply put; to have as many options as possible, but to strongly pursue only a few (hopefully one).

Given that totally determined and uniform probability distributions are the extreme cases for lowest and highest entropy values respectively, and considering that we'd like for the system-system and the environment-system cases to be as close as possible to said type of distributions; we can apply a distance information measure to see how far the actual repertoires are from the ideal ones, making use of the Earth Mover's Distance measure. We use this measure mainly for two reasons: first, other distance information measurements, such as KL-divergence (which was our first option, due its lack of geometrical symmetry, and its non-commutativity) or Mutual Information (for which there is a vast literature regarding causality) suffer too much from multiple zero values, something that can frequently occur in this context making them computationally unsuitable. Second, especially with respect to the comparison against ideal conditions, EMD displays a quite intuitive interpretation in terms of how much work, or energy, would be needed to change some particular repertoire into the desired probability distributions.

Being so, for the first environmental case ($E_{xy} = 2$) we obtain:

$$\delta_U = EMD(\, rep_{ey}(E_{xy} = 2) \mid\mid pU\,) = 0$$

Where pU stands for the discrete uniform distribution and δ_U for the difference to the ideal case (how far from total under-determination). Then, for the totally determined cases ($E_{xy} = 3$ and $E_{xy} = q$) we get:

$$\delta_U = EMD(\, rep_{ey}(E_{xy} = q) \mid\mid pU\,) = 0.5$$

Which, given the simplicity of the example at hand, can be interpreted directly like the ideal environment-system condition (insofar as eventually allowing some form of agency) and the total opposite scenario, respectively. As it can be seen, the lower the entropy value, the higher the chances for agency. Put another way, δ_U captures the lack of under-determination, or how (potentially) entangled the response of the system is with respect to the interpretation that it does from the environment.

Do note that, although the contrary is also true (the higher the distance value, the lower the space for the system to act on its own), the value of this metric is not something fixed ($\delta_U = 1$, for example) because, given the EMD algorithm, it depends significantly on two features, the number of possible elements and the distance among the elements within the distributions themselves. This is actually particularly relevant for the comparisons of the system-system repertoires, as the totally determined system-system future state will be represented by one state being one and the remaining being zero. Hence, the position of the value $p_i = 1$ (so the difference between that state and the rest of the possible states of the system) will matter.

We believe, however, that it is not necessary to incur an unfeasible number of calculations (like power-set permutations or something alike), but only to compare the repertoire against the totally determined case in which the highest probability within the repertoire is taken as the only possible (i.e.; $max(rep_{xy}(C_x = X)) = 1$, while all remaining elements are made zero), as this would represent the minimum value for such change to be possible and any other case will be therefore, less likely.

Accordingly, we will denote this construction as pF (focused) and δ_F as the difference (in work, or energy) required to transform the $rep_{xy}(C_x)$ into pF. For the system repertoires we obtain:

$$\delta_F = EMD(\, rep_{xy}(C_x = 0) \mid\mid pF\,) = 0.22$$

$$\delta_F = EMD(\, rep_{xy}(C_x = 1) \mid\mid pF\,) = 0.33$$

Which basically indicate the distances (i.e., amount of energy/work) that would be required to transform the state of the system into the ideal (the most focused) case possible. Unlike the previous case (δ_U), for δ_F the interpretation is a bit less straightforward, because it conceptually requires us to embrace the impossibility of mind-like phenomena. Otherwise, we would probably wish to compare also against a uniform distribution, along the lines of δ_U, expecting the *agent* to be capable of meaningfully acting by exerting as low causal power as possible (so to look for possible sparks of agency). This probably requires further elaboration that is out of the scope of the current work, however, we will briefly expand on this on the Discussion section.

Resuming with the previous comparison (δ_U), the closer to zero the values are, the higher the chances are for a system to actively determine (eventually, *purposely* undergo). This is because, given that the EMD is an unbounded metric (like most others in the case of information), we have chosen to make comparisons in terms of proximity to zero, with the intention to have a better intuition.

Finally, as we anticipated, we can also apply EMD to compare between the repertoires. Given that we know that for $E_{xy} = 3$ and $E_{xy} = q$ there's no possible under-determination, we will just examine the $E_{xy} = 2$ case:

$$\delta_{xe}(C_x = 0) = EMD(\, rep_{xy}(C_x = 0) \mid\mid rep_{ey}(E_{xy} = 2)\,) = 0.28$$

$$\delta_{xe}(C_x = 1) = EMD(\, rep_{xy}(C_x = 1) \mid\mid rep_{ey}(E_{xy} = 2)\,) = 0.17$$

Which, again, is simple to relate to previous results and our knowledge of the minimal dynamics of the isolated cell in the GoL. Here, as we would have expected from the above EMD comparisons, we see that the distance between the repertoire for $C_x = 1$ and $E_x = 2$ is less than for $C_x = 0$.

Whereas the numerical results we have seen until now are quite evident themselves, or at least seem to be easy to follow; this, however, will not always be the case as the complexity of the systems being examined increases. The purpose of this work has been to demonstrate the minimal possible case in the most intuitive fashion, so that we could develop an intuition about these methods. In the following section, we will explore a more complex case, from the dynamics of emergent patterns in the Game of Life. However, the scope remains the same and we will avoid unnecessary complications as much as possible.

4.3. A More Complex Case from GoL

Along the same lines of previous work, emergent patterns in GoL can be characterized as having different probabilities of maintaining their structural states, transitioning into new structures (i.e., instances of the same organization), or disintegrating. All of these are consequences of the spontaneous dynamics of their interaction with the environment as emergent units, which has proven to be a good ground for toy models of autopoiesis and autonomy [56,57,90–92]. Because of this, and in order to further test our intuitions, we analyzed transitions among a set of different GoL patterns and applied the same measurements that we used for a single cell (the patterns we will discuss are presented in Figure 2).

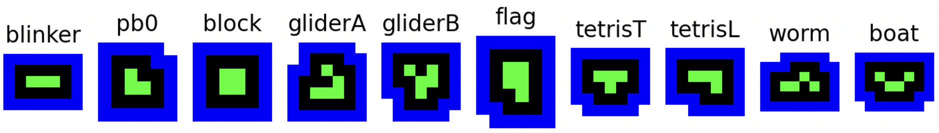
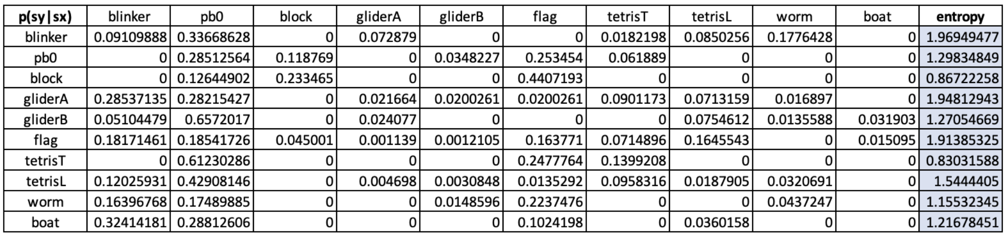


Figure 2. Some of the patterns of the GoL we investigated. Green represents active cells, black non-active cells acting as the 'membrane' of the system, whereas blue represent the environmental cells surrounding the system, which may be active or not.

To this end, we simulated all the possible transitions that all of these (and other) patterns could undergo (2^{20} to 2^{24}, depending on the number of environmental cells) and searched for the number of occurrences of the same patterns in the grid domain after transitions. Then we built up the system and environment repertoires by computing the probability distributions from the counts and, finally, measured the entropy of each individual distribution, while making comparisons between them using the EMD as we did for the single cell. First of all, given our knowledge of the transition counts, to obtain the system-system repertoires and entropy is a very much a direct process. The results from this are displayed in Figure 3.



p(sy|sx)	blinker	pb0	block	gliderA	gliderB	flag	tetrisT	tetrisL	worm	boat	**entropy**
blinker	0.09109888	0.33668628	0	0.072879	0	0	0.0182198	0.0850256	0.1776428	0	1.96949477
pb0	0	0.28512564	0.118769	0	0.0348227	0.253454	0.061889	0	0	0	1.29834849
block	0	0.12644902	0.233465	0	0	0.4407193	0	0	0	0	0.86722258
gliderA	0.28537135	0.28215427	0	0.021664	0.0200261	0.0200261	0.0901173	0.0713159	0.016897	0	1.94812943
gliderB	0.05104479	0.6572017	0	0.024077	0	0	0	0.0754612	0.0135588	0.031903	1.27054669
flag	0.18171461	0.18541726	0.045001	0.001139	0.0012105	0.163771	0.0714896	0.1645543	0	0.015095	1.91385325
tetrisT	0	0.61230286	0	0	0	0.2477764	0.1399208	0	0	0	0.83031588
tetrisL	0.12025931	0.42908146	0	0.004698	0.0030848	0.0135292	0.0958316	0.0187905	0.0320691	0	1.5444405
worm	0.16396768	0.17489885	0	0	0.0148596	0.2237476	0	0	0.0437247	0	1.15532345
boat	0.32414181	0.28812606	0	0	0	0.1024198	0	0.0360158	0	0	1.21678451

Figure 3. Entropy values obtained from the $S_x \to S_y$ repertoires, measuring system-system influence. H_x stands for $H(rep_{xy}(S_x))$, so for the entropy of such distributions.

After this, in order to obtain a measure of the effect of the environment over the system, we first, following the steps from the previous section, separated the possible state-environmental categories combinations, to then look for the subsets of these equivalent categories (the elements e_i within each E_{xy}) that could have been interpreted differently if the state of the system itself were to be different.

Thus, given some enaction $(S_x, e_i) \rightarrow S_y$, where $e_i \in E_{xy}$, we want to find at least a set $K_x \subset E_{xy}$, for which an arbitrary (viable) state S_u, under the same environmental circumstances e_i, could have produced another enaction: $(S_u, e_i) \rightarrow S_z$, where $S_u \neq S_z$ and $S_z \neq S_y$, and where the set $Alt(S_x) = \{K_x^1, ..., K_x^n\}$ describes all the alternative interpretations for each of the n states of the system.

In other words; because we know that, if there are no alternative transitions to some $S_x \rightarrow S_y$ given some external conditions, then S_y is being determined by the environmental perturbation on the system, inasmuch as it (the system) does not have any interpretation availability except for the one at hand. Therefore, we wish to find as many cases as possible, where the recursive dynamics of the system will trigger divergent transitions for identical environmental conditions (by classifying/interpreting them as an element of different equivalent categories). Hence, where identical objective environmental states are not locally equivalent and viceversa.

For instance, if we start by looking at the transitions between the 2 canonical states of the glider (denoted as 'gliderA' and 'gliderB' in Figure 2, gA and gB respectively from now on), we find that, for $gA \rightarrow gB$ there is just one alternative ($worm \rightarrow gB$) accounting only for 32 cases of a total of 32,800. Leading to $H(rep_{ey}(gA, alt(gB)) = 0.0112$.

Similarly, for the opposite state transition $gB \rightarrow gA$, we get: $H(rep_{ey}(gB, alt(gA)) = 0.0651$, which although higher due to more structural alternatives (including a recurrent case to gA ($n = 32$)), is still quite low compared to the entropy values of xy repertoires presented in Table 3. This, of course, is to be expected from a toy scenario such as GoL, especially taking into the account the low level of complexity of the patterns at hand.

Ideally, as we have discussed above, we would like to encounter an entropy value as close to zero as possible for the $S_x \rightarrow S_y$ repertoire (i.e., strongly focused) and, conversely, the highest possible for the entropy of the $E_x \rightarrow S_y$ repertoire, close to a uniform distribution, which would imply that the system has as many responses for a given environmental case, as viable states. With this in mind, we would examine cases along these lines.

Table 3. Comparison of entropy values obtained from the repertoires correspondent to the GoL patterns from Figure 2. S_x stands for the state of the system, while $HE(S_u)$ is short for $H(rep_{ey})(S_x, alt(S_u)))$, hence for the environment-system repertoires. The subindices $ifmax$ and min refer to the next maximum entropy value for each S_x (or the next after blinker and pb0, if lower) and minimum respectively (further details in the main text).

S_x	$HE_{alt}(blinker)$	$HE_{alt}(pb0)$	HE_{ifmax}	HE_{min}
blinker	0.633	1.322	tetrisL=0.381	gA=0.030
pb0	nf	1.583	block=0.389	gB=0.089
block	nf	0.570	flag=0.687	block=0.133
gliderA	0.235	1.251	tetrisL=0.151	gB=0.011
gliderB	0.767	1.340	tetrisL=0.266	gA=0.065
flag	0.644	0.949	gB=0.352	gA=0
tetrisT	2.060	nf	flag=0.591	tetrisT=0.525
tetrisL	1.352	1.853	flag=1.139	gA=0.120
worm	0.986	0.929	worm=0.524	gB=0.095
boat	1.445	1.314	flag=0.740	tetrisL=0.373

As it is visible from the results in Table 3, there is a general tendency towards lower values from environment-system entropy measures, along the lines we have discussed until now. As a matter of fact, the only case in which $H(rep_{ey}) > H(rep_{xy})$ (i.e., that the environment-system potential influence is lower than than that of the reciprocal system-system influence), is for the *tetrisL* pattern, when transitioning into a blinker.

Clearly this does not mean that this GoL emergent pattern can display agency in the cognitive connotation of the term, nor that it is choosing to undergo that specific state-transition, which would require representational and probably also phenomenological properties which are alien to minimal autonomous systems. Notwithstanding, the point that we would like to make here is that the possibility of choosing seems to exist, or at least, the formal development we have presented suggests so. And, being so, we believe that future work should be focused on understanding how cognitive systems can develop through the exploitation of this feature.

5. Discussion and Concluding Remarks

We have proposed a slightly different approach to the problem of agency in the context of enactive cognitive science [18,28,30,44]; that strictly in terms of causal influences, the continuous undergoing structural changes of a system and its environment are not uniform, so that the notion of system-environment co-determined behavior can be asymmetrical and dynamic.

Previously, in more autopoietic views of cognition [36,93], in the case of a coupled system-environment pair, it was asserted that physically there is a symmetry that autonomous systems are not capable of breaking (unless we invoke higher order cognitive processes). We have described here, though, how another kind of symmetry may be viable. Namely, the determination of the future state of the system itself, which does not require any form of mental representation whatsoever and for which we believe it is possible to characterize a non homogeneity of causal contributions. For this purpose, we have developed some formulations that can be applied in discrete cases, albeit we acknowledge that a more direct formulation instead of a series of comparisons would be a great advance.

In pure proto-cognitive terms (i.e., insofar as non mental, autonomous intelligent behavior), the specific self-persisting nature of autonomous systems would allow them to pass through multiple events of a higher causal contribution to the determination of its later state, than that of its environment, even if this is statistically uncommon as we saw in the latter case of GoL examples. This matches well with the proposed idea of *flickering* emergence [82], whereby a system may be conceived to be discontinuously emergent instead of in a binary emergent/non-emergent fashion. In the case we presented here, this would be translated as a protocognitive capacity, sporadically enabling self-determination, only when some (environmental and systemic) conditions can be met.

In this sense, the key difference between environment and (autonomous) system, is that the latter remains as such, instead of basically dissolving into something new. In particular, we suggest the origin of agency, insofar as a minimal property enabling under-determination, can be traced (logically and evolutionarily) to the mechanistic dynamics of autonomous organizations, hence prior to biological or mental phenomena. Do note, however, that underpinning this organizational stability must be a high degree of coercion from the system over the local elements and their responses; otherwise the selectivity of the system would be unreliable and inconsistent, as well as any kind of mapping. Autonomous organizations, in this respect, operate as subordinating mechanisms by which the otherwise much wider domain of responses of every component becomes narrowed in order to produce coherent global transitions (autopoiesis being the strongest case in this regard) [42,94]. Such local-to-global subordination, however, as we have seen, doesn't necessarily entail a narrow set of responses for the system as a whole, as the degree of complexity of its organization is not dependent on the sum of the possible responses of every element, but on the number of valid mappings (arising from the interrelated global responses) which permit valid (viable and consistent) transitions between any two valid states. To some extent this resembles related work, which proposed mapping interpretations in terms of bayesian reasoning [33,34]; we think this could be a productive line of future work.

A final important point we should address is the two-folded mapping we have developed. In this respect, we believe that the first half of the overall idea is more robust than the second. More specifically, while the underdetermination as a consequence of an expanded set of available interpretations of environmental perturbations, by which the state of the environment becomes the most important factor, can be directly derived from system-environmental dynamics, the purported requirement for its exploitation, that is, the statistical concentration of its potential responses into one, may be counterproductive when considered in the context of later exploitation by more complex systems.

Put another way, it could perfectly be the case that, the opposite case (i.e., a highly underdetermined output) could be easier to exploit.

Along the same lines of protocognitive capacities, interesting further work could involve the exploration of this same phenomena in small groups of minimal systems, to see, first, whether they are capable of more effective environmental influence regulation and, more importantly, to explore the possibility of some form of collective exploitation of the individual underdetermined mappings, whereby the actual underdetermined system would be the group.

Summarizing, we have presented the idea that, while enaction is a single process, it involves different components and it has an internal structure that can be characterized as a mapping. Being so, a more complex mapping can be seen as disentanglement or decoupling from an otherwise straightforward one-to-one interpretation, because it can provide more than one action, given the same interpretation or viceversa (insofar as this relation is given by equivalent categories).

For every protocognitive property that we have explored so far, we have used the Game of Life as our ground for testing and building proof of concepts. This strategy, however, has a fundamental limitation that may have become evident as this point, namely, that the dynamics of the state-transitions are too constrictive. Do note that the bottom of the issue goes beyond the simultaneous update of the whole grid, but deeper, to the implicit idea that cognitive (and protocognitive) properties somehow have to follow the temporal logic that enables them.

Funding: This research received no external funding.

Data Availability Statement: The datasets presented in this article are not readily available due to technical limitations (large size of respective files). Requests to access the datasets should be directed to f.rodriguez-vergara@sussex.ac.uk. The original code for the datasets presented in the study are openly available at: https://github.com/dinosaurioinvisible/gol

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Baltieri, M.; Iizuka, H.; Witkowsi, O.; Sinapayen, L.; Suzuki, K. Hybrid Life: Integrating biological, artificial, and cognitive systems. *WIREs Cognitive Science* **2023**, p. e1662. doi:https://doi.org/10.1002/wcs.1662.
2. Tallis, R. *Freedom: An impossible reality*; Agenda Publishing Limited, 2021.
3. Libet, B.; Gleason, C.A.; Right, E.W.; Pearl, D.K. TIME OF CONSCIOUS INTENTION TO ACT IN RELATION TO ONSET OF CEREBRAL ACTIVITY (READINESS-POTENTIAL): THE UNCONSCIOUS INITIATION OF A FREELY VOLUNTARY ACT. *Brain* **1983**, *106*, 623–642. doi:https://doi.org/10.1093/brain/106.3.623.
4. Kim, J. *Mind in a physical world: an essay on the mind-problem and mental causation*; MIT Press, 1998.
5. Churchland, P.; Suhler, C. Agency and Control: The Subcortical Role in Good Decisions. In *Moral Psychology, Volume 4: Free Will and Moral Responsibility, Walter Sinnott-Armstrong.*; The MIT Press, 2014.
6. Dennet, D. *Elbow Room: The Varieties of Free Will Worth Wanting*; The MIT Press, 2015.
7. Hill, T.T. Neurocognitive free will. *Proceedings of the Royal Society B* **2019**, *286*, 20190510. doi:http://dx.doi.org/10.1098/rspb.2019.0510.
8. Lavazza, A. Why Cognitive Sciences Do Not Prove That Free Will Is an Epiphenomenon. *Frontiers in Psychology* **2019**, *10*. doi:https://doi.org/10.3389/fpsyg.2019.00326.
9. Abramova, K.; Villalobos, M. The apparent (Ur-)Intentionality of Living Beings and the Game of Content. *Philosophia* **2015**, *43*, 651–668.
10. Potter, H.; Mitchell, K. Naturalasing agent causation. *Entropy* **2022**, *24*. doi:https://doi.org/10.3390/e24040472.
11. Barandiaran, X.; Di Paolo, E.; Rohde, M. Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality in Action. *Adaptive Behavior* **2009**, *17*, 367–386. doi:10.1177/1059712309343819.
12. Seifert, G.; Sealander, A.; Marzen, S.; Levin, M. From reinforcement learning to agency: Frameworks for understanding basal cognition. *BioSystems* **2024**, *235*, 105107. doi:https://doi.org/10.1016/j.biosystems.2023.105107.
13. Moreno, A.; Etxeberria, A. Agency in Natural and Artificial Systems. *Artificial Life* **2005**, *11*, 161–175.
14. Biehl, M.; Virgo, N. Bayesian ghosts in a machine? ALIFE 2023: Ghost in the Machine: Proceedings of the 2023 Artificial Life Conference, 2023, p. 96. doi:https://doi.org/10.1162/isal_a_00607.

15. Rovelli, C. Agency in Physics. *arXiv:2007.05300v2* **2020**. doi:https://doi.org/10.48550/arXiv.2007.05300.
16. Yuan, B.; Zhang, J.; Lyu, A.; Wu, J.; Wang, Z.; Yang, M.; Liu, K.; 2, M.M.; Cui, P. Emergence and Causality in Complex Systems: A Survey of Causal Emergence and Related Quantitative Studies. *Entropy* **2024**, *26*. doi:https://doi.org/10.3390/e26020108.
17. Biehl, M.; Virgo, N. Interpreting systems as solving POMDPs: a step towards a formal understanding of agency. In *Buckley, C.L., et al. Active Inference. IWAI 2022, Communications in Computer and Information Science, vol 1721.*; Springer, Cham, 2023. doi:https://doi.org/10.1007/978-3-031-28719-0_2.
18. Froese, T. Irruption Theory: A Novel Conceptualization of the Enactive Account of Motivated Activity. *Entropy* **2023**, *25*, 748.
19. Maturana, H.; Varela, F. *Autopoiesis: the organization of the living. [De maquinas y seres vivos. Autopoiesis: la organizacion de lo vivo]. 7th edition from 1994.*; Editorial Universitaria, 1973.
20. Maturana, H. Tha organization of the living: A theory of the living organization. *International Journal of Man-Machine Studies* **1975**, *7*, 313–332. doi:https://doi.org/10.1016/S0020-7373(75)80015-0.
21. Varela, F. *Principles of Biological Autonomy*; North Holland, 1979.
22. Varela, F. Patterns of life: Intertwining identity and cognition. *Brain cognition* **1997**, *34*, 72–87.
23. Varela, F.; Thompson, E.; Rosch, E. *The embodied mind: Cognitive science and human experience*; The MIT Press, 1991.
24. Varela, F., Preface from Francisco J. Varela Garcia to the second edition. In *De Mâquinas y seres vivos. Autopoiesis: la organizaciôn de lo vivo*; Editorial Universitaria, 1994; chapter Preface.
25. Di Paolo, E. Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences* **2005**, *4*, 429–452. doi:10.1007/s11097-005-9002-y.
26. Barandiaran, X. Autonomy and Enactivism: Towards a Theory of Sensorimotor Autonomous Agency. *Topoi* **2017**, *36*, 409–430. doi:10.1007/s11245-016-9365-4.
27. Beer, R. An integrated Perspective on the Constitutive and Interactive Dimensions of Autonomy. Proceedings of the ALIFE 2020: The 2020 Conference on Artificial Life, 2020, pp. 202–209. doi:10.1162/isal_a_00245.
28. Ward, M.; Silverman, D.; Villalobos, M. Introduction: The Varieties of Enactivism. *Topoi* **2017**, *36*, 365–375.
29. Gallagher, S. *Embodied and Enactive approaches to Cognition*; Cambridge University Press, 2023.
30. Buhrmann, T.; Di Paolo, E. The sense of agency - a phenomenological consequence of enacting sensorimotor schemes. *Phenomenology and the Cognitive Sciences* **2017**, *16*, 207–236.
31. Dennett, D.C. Intentional Systems. *The Journal of Philosophy* **1971**, *68*, 87–106.
32. Dennet, D.C. The intentional stance in theory and practice. In *R. W. Byrne & A. Whiten (Eds.), Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans*; Clarendon Press/Oxford University Press., 1988.
33. Biehl, M.; Kanai, R. Dynamics of a Bayesian Hyperparameter in a Markov Chain. In *Verbelen, T., Lanillos, P., Buckley, C.L., De Boom, C. (eds) Active Inference. IWAI 2020. Communications in Computer and Information Science, vol 1326*; Springer, Cham, 2020.
34. Virgo, N.; Biehl, M.; McGregor, S. Interpreting Dynamical Systems as Bayesian Reasoners. In *Kamp, M., et al. Machine Learning and Principles and Practice of Knowledge Discovery in Databases. ECML PKDD 2021. Communications in Computer and Information Science, vol 1524*; Springer, Cham, 2021. doi:https://doi.org/10.1007/978-3-030-93736-2_52.
35. Hutto, D.; Myin, E. *Radicalizing Enactivism. Basic minds without content.*; MIT Press, 2012.
36. Villalobos, M.; Silverman, D. Extended functionalism, radical enactivism and the autopoietic theory of cognition: prospects for a full revolution in cognitive science. *Phenomenology and the Cognitive Sciences* **2018**, *17*, 719–739.
37. Varela, F. Two Principles for Self-Organization. In *Ulrich, H., Probst, G.J.B. (eds.) Self-Organization and Management of Social Systems*; Springer Series on Synergetics, vol 26. Springer, Berlin, Heidelberg., 1984.
38. Virgo, N.; Harvey, I. Adaptive growth processes: a model inspired by Pask's ear. Artificial Life XI, 2008.
39. Sayama, H. Construction theory, self-replication, and the halting problem. *Complexity* **2008**, *13*, 16–22. doi:https://doi.org/10.1002/cplx.20218.
40. Hanczyz, M.; Ikegami, T. Chemical basis for minimal cognition. *Artificial Life* **2010**, *16*, 233–243.
41. Beer, R. Bittorio revisited: Structural coupling in the Game of Life. *Adaptive Behavior* **2020**, *28*, 197–212.
42. Villalobos, M.; Ward, D. Living Systems: Autonomy, Autopoiesis and Enaction. *Philosophy & Technology* **2015**, *28*, 225–239.

43. Hutto, D.; Myin, E. *Evolving Enactivism. Basic Minds Meet Content*; MIT Press, 2017.

44. Di Paolo, E.; Burhmann, T.; Barandarian, X. *Sensorimotor Life: An enactive proposal*; Oxford University Press, 2017.

45. Friston, K. Am I Self-Conscious? (Or Does Self-Organization Entail Self-Consciousness?). *Frontiers in Psychology* **2018**, *9*.

46. Bowes, S. *Naturally Minded: Mental Causation, Virtual Machines, and Maps*; Springer Verlag, 2023.

47. Froese, T.; Di Paolo, E. The enactive approach. Theoretical sketches from cell to society. *Pragmatics and Cognition* **2011**, *19*, 21–36.

48. Albantakis, L.; Barbosa, L.; Findlay, G.; Grasso, M.; Haun, A.M.; Marshall, W.; Mayner, W.G.P.; Zaeemzadeh, A.; Boly, M.; Juel, B.E.; Sasai, S.; Fujii, K.; Isaac David, J.H.; Lang, J.P.; Tononi, G. Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLoS Computational Biology* **2023**, *19*, e1011465. doi:https://doi.org/10.1371/journal.pcbi.1011465.

49. Chalmers, D. Strong and Weak Emergence. *In: The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion. Oxford University Press.* **2011**, pp. 244–256.

50. Dennett, D. Autonomy, Consciousness, and Freedom. *The Amherst Lecture in Philosophy* **2019**, *14*, 1–22.

51. Froese, T.; Taguchi., S. The Problem of Meaning in AI and Robotics: Still with Us after All These Years. *Philosophies* **2019**, *4*.

52. Cea, I. On motivating irruptions: the need for a multilevel approach at the interface between life and mind. *Adaptive Behavior* **2023**, *32*, 95–99. doi:10.1177/10597123231184651.

53. Froese, T. To Understand the Origin of Life We Must First Understand the Role of Normativity. *Biosemiotics* **2021**, *14*, 657–663. doi:https://doi.org/10.1007/s12304-021-09467-3.

54. Maturana, H., Preface from Humberto Maturana Romesîn to the second edition. In *De Mâquinas y seres vivos. Autopoiesis: la organizaciôn de lo vivo*; Editorial Universitaria, 1994; chapter Preface.

55. Maturana, H. Autopoiesis, Structural Coupling and Cognition: A history of these and othe notions in the biology of cognition. *Cybernetics and Human Knowing* **2002**, *9*, 5–34.

56. Beer, R. Autopoiesis and Cognition in the Game of Life. *Artificial Life* **2004**, *10*, 309–326.

57. Beer, R. The Cognitive Domain of Glider in the Game of Life. *Artificial Life* **2014**, *20*, 183–206.

58. Dell, P. Understanding Bateson and Maturana: Toward a Biological Foundation for The Social Sciences. *Journal of Marital and Family Therapy* **1985**, *11*, 1–20.

59. Ashby, W. *An introduction to cybernetics*; J. Wiley, New York, 1956.

60. Maturana, H. Everything said is said by an observer. In *Thompson W. I. (ed.) Gaia: A way of knowing*; Lindisfarne Press, New York, 1987.

61. Beeson, I. Implications of the Theory of Autopoiesis for the discipline and practice of Information Systems. In *Russo, N.L., Fitzgerald, B. De Gross, J.I. (eds) Realining Research and Practice in Information Systems Development. IFIP - The international Federation for Information Processing, vol. 66*; Springer, Boston, MA, 2017.

62. Seth, A. *Being you: A new science of consciousness*; Faber and Faber Ltd, 2021.

63. Bateson, G. *Steps to an echology of mind: Collected essays in anthropology, psychiatry, evolution and epistemology*; Jason Aronson, 1972.

64. Tononi, G. An Information Integration Theory of Consciousness. *BMC Neuroscience* **2004**, *5*.

65. Balduzzi, D.; Tononi, G. Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework. *PLoS Comput Biol.* **2008**, *4*, e1000091. doi:10.1371/journal.pcbi.1000091.

66. Oizumi, M.; Albantakis, L.; Tononi, G. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLOS Computational Biology* **2014**, *10*. doi:https://doi.org/10.1371/journal.pcbi.1003588.

67. Tononi, G.; Koch, C. Consciousness: here, there and everywhere? *Philosophical Transactions of the Royal Society B* **2015**, *370*. doi:https://doi.org/10.1098/rstb.2014.0167.

68. Pautz, A. What is the Integrated Information Theory of Consciousness. A Catalogue of Questions. *Journal of Consciousness Studies* **2019**, *26*, 188–215.

69. Yuan, B.; Zhang, J.; Lyu, A.; Wu, J.; Wang, Z.; Yang, M.; Liu, K.; Mou, M.; Cui, P. Emergence and Causality in Complex Systems: A Survey of Causal Emergence and Related Quantitative Studies. *Entropy* **2024**, *26*. doi:https://doi.org/10.3390/e26020108.

70. Lombardi, O.; López, C. What does 'Information' Mean in Integrated Information Theory. *Entropy* **2018**, *20*, 894.

71. Mediano, P.; Seth, A.; Barret, A. Measuring Integrated Information: Comparison of Candidate Measures in Theory and Simulation. *Entropy* **2018**, *21*, 17.

72. Chalmers, D. The Combination Problem for Panpsychism. In *Brüntrup Godehard & Jaskolla Ludwig (eds.), Panpsychism*; Oxford University Press, 2017.
73. Tsuchiya, N.; Taguchi, S.; Saigo, H. Using category theory to asses the relationship between consciousness and integrated information theory. *Neuroscience Research* **2016**, *107*, 1–7.
74. Doerig, A.; Schurger, A.; Hess, K.; Herzog, M. The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and Cognition* **2019**, *72*, 49–59.
75. Merker, B.; Williford, K.; Rudrauf, D. The Integrated Information Theory of consciousness: A case of mistaken identity. *Behavioral and Brain Sciences* **2021**, *May*, 1–72.
76. Singhal, I.; Mudumba, R.; Srinivasan, N. In search of lost time: Integrated information theory needs constraints from temporal phenomenology. *Philosophy and the Mind Sciences* **2022**, *3*. doi:https://doi.org/10.33735/phimisci.2022.9438.
77. Northoff, G.; Zilio, F. From Shorter to Longer Timescales: Converging Integrated Information Theory (IIT) with the Temporo-Spatial Theory of Consciousness (TTC). *Entropy* **2022**, *24*. doi:https://doi.org/10.3390/ e24020270.
78. Rodriguez, F.; Husbands, P.; Ghosh, A.; White, B. Frame by frame? A contrasting research framework for time experience. ALIFE 2023: Ghost in the Machine: Proceedings of the 2023 Artificial Life Conference. MIT Press, 2023, p. 75. doi:10.1162/isal_a_00688.
79. Aguilera, M.; Di Paolo, E. Integrated information in the thermodynamic limit. *Neural Networks* **2019**, *114*, 136–149.
80. Mediano, P.; Rosas, F.; Bor, D.; Seth, A.; Barret, A. The strength of weak integrated information theory. *Trends on Cognitive Sciences* **2022**, *26*, 646–655.
81. De Rosas, F.; Mediano, P.; Jensen, H.; Seth, A.; Barret, A.; Carthart-Harris, R. Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLOS. Computational Biology* **2020**, *16*.
82. Varley, T.F. Flickering Emergences: The Question of Locality in Information-Theoretic Approaches to Emergence. *Entropy* **2022**, *25*. doi:https://doi.org/10.3390/e25010054.
83. Rubner, Y.; Tomasi, C.; Guibas, J. A Metric for Distributions with Applications to Image Databases. Proceedings of the 1998 IEEE International Conference on Computer Vision, Bombay, India, 1998.
84. Weng, L. What is Wasserstein distance? https:/lilianweng.github.io/posts/2017-08-20-gan/what-is-wasserstein-distance.
85. Weisstein, E. Moore Neighborhood. From MathWorld–A Wolfram Web Resource. https://mathworld.wolfram.com/MooreNeighborhood.html.
86. Gardner, M. Mathematical Games: The Fantastic Combinations of John Conway's New Solitaire Game 'Life'. *Scientific American* **1970**, *223*, 120–123.
87. Berlekamp, E.; Conway, J.; Guy, R. *Winning ways for your mathematical plays, vol. 2*; New York: Academic Press, 1982.
88. Weber, A.; Varela, F. Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences* **2002**, *1*, 97–125.
89. Kirchhoff, M. Autopoiesis, free energy, and the life-mind continuity thesis. *Synthese* **2018**, *195*, 2519–2540.
90. Beer, R. Characterizing Autopoiesis in the Game of Life. *Artificial Life* **2015**, *21*, 1–19.
91. Beer, R. On the Origins of Gliders. Proceedings of the ALIFE 2018: The 2018 Conference on Artificial Life. ALIFE2018: The 2018 Conference on Artificial Life. Tokyo, Japan., 2018, pp. 67–74. doi:https://doi.org/10.1162/isal_a_00019.
92. Rodriguez, F.; Husbands, P. A saucerful of secrets: Open-ended organizational closure in the Game of Life. ALIFE 2024: Proceedings of the 2024 Artificial Life Conference. MIT Press, 2024, p. 4. doi:10.1162/isal_a_00712.
93. Villalobos, M.; Palacios, S. Autopoietic theory, enactivism, and their incommensurable marks of the cognitive. *Synthese* **2021**, *198*, 571–587. doi:https://doi.org/10.1007/s11229-019-02376-6.
94. Lyon, P. Autopoiesis and Knowing: Reflections on Maturana's Biogenic Explanation of Cognition. *Cybernetics And Human Knowing* **2004**, *11*, 21–46.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.