The enactive account of motivated activity and the hard problem of efficacy (HPE): Artificial life meets the physics of life

Tom Froese and Georgii Karelin

Embodied Cognitive Science Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa 904-0495, Japan
tom.froese@oist.jp

Abstract

The enactive approach to cognitive science has undergone a bio-phenomenologically inspired "normative turn" by characterizing an organism's activity as motivated by intrinsic value, where this value is grounded in adaptive self-production under precarious conditions. However, efforts in the field of artificial life to model this enactive conception of life have unwittingly revealed a case of what can be called *the hard problem of efficacy (HPE)*: how could any intrinsic value *as such* make an effective difference to an organism's behavior, in particular if bodily activity is purely determined by valueless material-organizational factors? First, this theoretical challenge of the HPE is formulated in the context of the enactive account of motivated activity. Then, by critically analyzing Schrödinger's work on the methodological principles that define the scientific world image, it is argued that they can be revised to allow solutions to the HPE. This involves placing a limit on Schrödinger's principle of understandability. The key move is to operationalize this limit with the concept of *irruption*: an organism's motivations can make a physical difference to its bodily activity, but only indeterminately so, akin to a breakdown of its material-organizational constraints. Irruptions can thereby indirectly facilitate behavior-switching as well as long-term self-organization of adaptive behavior. Finally, it is proposed that the efficacy of motivated activity has its own specific energy cost due to the disordering effect of irruptions, which provides a new perspective on agency and the notion of mental work.

Introduction

Since its inauguration in the 1990s, the enactive approach to cognitive science has been promoting the artificial life (Alife) route to artificial intelligence (AI) (e.g., Varela, 1995). It drew inspiration from the tradition of modeling autopoiesis (Varela, Maturana, & Uribe, 1974) to highlight the notion of biological autonomy (Bourgine & Varela, 1992): a living system is said to constitute its own identity, and hence thereby give rise to an intrinsic perspective on the significance of its environment for the maintenance of that identity. Complex organisms are an integration of various such autonomous systemic identities, including of neural, immune, and sensorimotor systems, that are integrated into a meshwork of emergent or selfless selves (Varela, 1991). The enactive approach has since continued to develop these ideas, especially in terms of the concept of sense-making (Di Paolo, Buhrmann, & Barandiaran, 2017).

This generalized notion of constitutive autonomy has been useful for the ALife route to AI, because it allowed cognitive and evolutionary robotics researchers to shift their focus from the biological autonomy of *life itself* to that of *ways of life* (Di Paolo, 2003, 2010). It is difficult to scale up the conditions for the material self-production of metabolic networks to the level of agent-based simulation models of "cognitively interesting behavior" (Beer, 1997), but scaling up the dynamics of habitual sensorimotor patterns is presumably more feasible (Egbert & Barandiaran, 2014; Ramírez-Vizcaya & Froese, 2020).

However, despite interest in recreating the conditions for the emergence of artificial agency with an intrinsic perspective, it is not clear whether this enactive approach to artificial life has managed to solve AI's classic problems, especially the problem of meaning – how symbolic or sub-symbolic data could appear as meaningful from the perspective of the artificial system itself (Dreyfus, 2007). Froese and Ziemke (2009) had identified two conditions for an enactive AI, that is, for a synthetic realization of sense-making: constitutive autonomy and adaptivity. It has been generally easier to satisfy each of these conditions apart from the other, depending on whether the synthetic approaches focused on minimal self-production or minimal cognition. But by now there are artificial life models that have managed to come close to satisfying both of these conditions (e.g., Agmon, Gates, & Beer, 2016). However, these advances have also made it evident that nothing qualitatively different happens when both conditions are satisfied, at least not in simulation models. Although the model now includes a simulated living system whose activity is potentially more interesting to study, because it is simulating both self-producing and adaptive processes, the scope for an intrinsic perspective to play any role in behavior in its own right is ruled out; all behavior is predetermined by the simulation model's initial conditions and dynamical laws.

What happened to the enactive approach's intriguing idea that such an artificial autonomous-adaptive system could act differently based on its intrinsically meaningful perspective? Unfortunately, the deflationary conclusion seems to be that, even if the enactive approach is on the right track and sense-making were to emerge, the resulting meaning cannot be the difference that actually makes a difference. A critical analysis by Froese and Taguchi (2019) revealed that the classic problem of meaning is therefore still very much a concern for enactive AI: for as long as the artificial agent's behavior is determined by the agent-environment system's dynamics, any meaning that were to somehow arise for the agent's intrinsic perspective simply cannot make a difference in its own right to the agent's behavior. Froese and Taguchi therefore concluded that, for the enactive approach to make conceptual room for an efficacious sense-making, this requires rejection of the assumption of state-determinism.

But the rejection of determinism would only be a first step: it would only open the logical possibility of an agent behaving in a manner that is not completely predetermined by the laws governing the system's states. In itself, this theoretical move does not make it intelligible how any other factors, such as the meaning something has for an agent, could make a difference in its own right in the generation of its behavior, and to do so in a way that is in accordance with that meaning. It seems that enactive AI will also need to face up to another classic problem in cognitive science, which has sometimes been debated under the heading of mental causation (Hanna & Maiese, 2009).

These considerations widen the scope of the diagnosis: the impasse faced by enactive AI is matched by an impasse of the enactive conception of life more generally. First, its emphasis on the self-constitution and maintenance of a system's identity as the root of normativity gives it a rather conservative scope compared to the open-endedness of life and mind (Barbaras, 2010; Barrett, 2017). Moreover, the current enactive account of sense-making leaves it unclear how an agent's value or other motivations *as such*, that is, as an intrinsic perspective of sense-making imbued with significance, meaning, and purpose, could in principle make a difference in its own right to its behavior. In the simulations, for example, activity can be fully described in terms of abstract dynamics (Barandiaran & Egbert, 2014), but thereby leaving impotent any observer-relative attribution of value or other motivations. The situation of a real organism's motivated activity may be a different matter, but the differences between simulation and reality tend to remain unspecified and their relevance for overcoming this impotence unclear. We will refer to this specific problem of how an organism's motivations as such can make a difference in their own right to its activity as the *hard problem of efficacy* (HPE). Arguably, most efforts have focused on trying to explain how living matter can also be a mind (Thompson, 2007), consciousness (Varela, 1996), or contentful (Hutto & Myin, 2017). The HPE refers to the other direction of influence: assuming that an organism's activity is motivated, for example by value, meaning, mental content, or by full-blown consciousness, how could such factors make a difference in their own right to the organism's material basis?

Moreover, the enactive approach has kept its distance from existing attempts at solving the HPE, which puts it in a strange middle position. For example, although the enactive approach rejects representationalism, critics have argued that some of its core concepts, such as adaptivity, sense-making, and know-how, have affinities with representational accounts (Hutto & Myin, 2013; Villalobos, 2013). Similarly, although the enactive approach adopts a naturalist orientation, it is at the same time oddly disinterested in engaging with advances in the physics of life, such as appealing to the possibility that organisms' tend to maximize entropy production as a way to ground normativity in a more open-ended manner (Barrett, 2020a).

What this equivocal explanatory stance seems to suggest is that the enactive approach is struggling with how to integrate two opposing commitments regarding its conception of life, namely, to allow that both an agent's *intrinsic perspective* and its *external embodiment* jointly make a difference to behavior. A detailed assessment of this apparent tension has already been provided elsewhere (Froese, 2023), so the following will be a brief recap to set the scene for the rest of this analysis.

This enactive version of the traditional mind-body problem is sometimes discussed in phenomenological terms using the concept of the "lived body" (Hanna & Thompson, 2003), which is posited to be the interface between the material body and the mental self, with the aim of turning the dualistic mind-body problem into a more tractable triadic mind-body-body problem. However, while this theoretical move displaces the mind-body problem, it ultimately does not manage resolve it: instead, it raises the question of how this phenomenological notion of the lived body could become integrated into the hard sciences of life and mind, and more specifically in such a way that it can become intelligible how an agent's lived experience makes a difference to its bodily material processes. Solving this version of the HPE becomes even more pressing given that there are related approaches, such as the classic autopoietic theory of Maturana's biology of cognition, that try to sidestep the HPE altogether by rejecting teleological explanations (Villalobos & Ward, 2016).

The mind-body problem remains sufficiently daunting that some branches of enactivism have opted to give up on solving the HPE: they deny the possibility that our lived experience, and agential motivations more generally, can have their own effect on behavior. For example, they are instead developing an embodied-enactive take on mind-brain identity theory (Myin & Zahnoun, 2018). Yet, like in the case of Maturana's biology of cognition, accepting identity theory borders on the acceptance of epiphenomenalism – our lived experience would become an impotent quirk of nature.

This odd implication of the inefficacy of lived experience, and of agency more generally, is rejected as implausible not only by enactivism's original neurophenomenology research program (Varela & Thompson, 2003), but also increasingly by more mainstream approaches to the science of consciousness (e.g., Cleeremans & Tallon-Baudry, 2022). Importantly, giving up on solving the HPE is in direct tension with the enactive approach's founding commitment to bringing human lived experience closer to cognitive science (Varela, Thompson, & Rosch, 2017), which in turn is largely dependent on securing the possibility that our intrinsically meaningful perspective on the world has efficacy in its own right (T. Fuchs, 2021).

To make progress on the deep theoretical challenge of the HPE, it helps to make explicit the two opposing conditions that the enactive approach is committed to satisfying. Hence, we propose the following scientific mission statement: *the enactive approach aims to account for the motivation- and matter-involving generation of motivated activity*. In the next section we will analyze the scientific principles that stand in the way of resolving the tension of this double commitment.

The principle of objectification

In a nutshell, the ambition to make conceptual room for the involvement of an agent's motivations in the generation of its behavior entails that the standard premise of determinism needs to be rejected (Froese & Taguchi, 2019). This entailment in turn highlights an interesting affinity between the enactive approach to cognitive science and a libertarian philosophy of human freedom (T. Fuchs, 2021; Gallagher, 2017; Tewes, 2017), the latter of which explicitly rejects the possibility of a clockwork or block universe.

However, the mere replacement of a deterministic concept of nature with a non-deterministic one does not go far enough in clarifying precisely how motivations and matter can both be involved in the co-dependent generation of an agent's behavior. Accordingly, there has been growing recognition by enactive and phenomenological approaches that the next substantial advances in rethinking the mind must ultimately go hand-in-hand with rethinking "nature" (Gallagher, 2018; Hutto & Satne, 2018b; Zahavi, 2017). For this purpose, it is helpful to examine some of the main roots of the scientific method, on which our modern concept of nature is based.

Fortunately, we can gain a useful overview of what is at stake from Schrödinger's 1956 lectures on *Mind and Matter*, which analyzed the mind-body problem in the context of the basic scientific principles that constitute the scientific world image. Previously, he had traced these scientific principles to more ancient conceptions of nature in his 1948 lectures on *Nature and the Greeks* (Schrödinger, 1992, 1996). In these lecture series Schrödinger proposed that it is not an accident that there is no room for our motivational involvement in bodily activity in the domain of nature. Rather, this is a direct consequence of what he came to call the "principle of objectification", and which he described as follows:

> I maintain that it amounts to a certain simplification which we adopt in order to master the infinitely intricate problem of nature. Without being aware of it and without being rigorously systematic about it, we exclude the Subject of Cognizance from the domain of nature that we endeavor to understand. We step with our own person back into the part of an onlooker who does not belong to the world, which by this very procedure becomes an objective world. (Schrödinger, 1992, p. 118)

However, Schrödinger highlights that the scientific domain of nature derived by this principle of objectification faces the challenge that everyone of us is included in that nature: every subject is also objectified, i.e., into nothing but a body-object that is contained in the world like any other physical object. It is tempting for us to therefore conclude that "I myself also form part of this real material world around me. I so to speak put my own sentient self (which had constructed this world as a mental product) back into it" (ibid., p. 119) – a conclusion that comes close to the kind of reasoning that motivates the adoption of the representationalist framework by contemporary mainstream cognitive science. However, the principle of objectification is not as easily reversed as the representationalists would like to believe, and Schrödinger warned of "the pandemonium of disastrous logical consequences that flow from the aforesaid chain of faulty conclusions" (ibid., p. 119).

Importantly, we need to avoid committing the fallacy of looking for our subjective first-person perspective inside of a scientific construct that was enabled by removing all subjective perspective in the first place. This does not mean that the "Subject of Cognizance" is not part of reality as a whole; to the contrary, it also has a place in our conception of reality, as it is what enables the scientific world image in the first place. There is therefore no contradiction between accepting the reality of the physical world and the reality of subjective experience (van Buuren, 2018). Rather, the concern is methodological: Given that the principle of objectification is always already at work in scientific observation, it has the unavoidable consequence of casting our experiential lifeworld, including our lived bodies, literally into an external object of study, namely the universe as the scientific domain of nature as studied by physics. Hence, we must take care to maintain the special conceptual hygiene demanded of us as observers when casting all reality into an object of study for natural science:

> The material world has only been constructed at the price of taking the self, that is, mind, out of it, removing it; mind is not part of it; obviously, therefore, it can neither act on it nor be acted on by any of its parts. (Schrödinger, 1992, p. 119)

Schrödinger brings us remarkably close to the methodological source of two open fundamental problems faced by cognitive science, which are also recognized by enactivism: the hard problem of consciousness (Varela, 1996), and the hard problem of mental content more generally (Hutto & Myin, 2013). He also prefigures the HPE – on this view, the observer's mind is seen as powerless to act on the material world. Basically, the principle of objectification as a condition of possibility of turning the domain of nature into a target of physics amounts to what Schrödinger refers to as an "exclusion principle":

> the conscious mind itself remains a stranger within that construct, it has no living space in it, you can spot it nowhere in that space. We do not usually realize this fact, because we have entirely taken to thinking of the personality of a human being, or for that matter also that of an animal, as located in the interior of its body. To learn that it cannot really be found there is so amazing that it meets with doubt and hesitation, we are very loath to admit it. (Schrödinger, 1992, p. 122)

For Schrödinger, the solution demanded by this impasse would involve removing the methodological principles that gives to rise to the exclusion principle, but he also admits that this does not readily translate into a practically workable solution: to give up on the principle of objectification would essentially require rebuilding science anew, and he was concerned that there is not yet sufficient knowledge to vindicate such a change in scientific practices that otherwise worked for centuries. Yet even if the principle of objectification cannot be avoided, there is another principle at the heart of the scientific image that may be more malleable for finding solutions to the HPE.

The principle of understandability

Schrödinger considered interpretations of quantum physics by his contemporaries that were supposed to allow for a direct causal influence of the subject on the object, but he did not find a compelling argument against the exclusion principle based on the principle of objectification. Still, Schrödinger was open to the possibility that the apparent strangeness of nature that was revealed by quantum physics, specifically the uncertainty principle, is indicative of a historically first breach of another fundamental principle of the scientific method, namely, of what he called the "principle of understandability." In contrast to the deterministic picture of nature leading up to quantum physics, it is now fairly uncontroversial to claim that it is consistent with all that we know from physics that there are happenings in the universe for which no fully deterministic physical description can ever be provided (Conway & Kochen, 2009).

Surprisingly, Schrödinger did not connect his discussion of the potential discovery of fundamental limits to the principle of understandability back to his preceding reflections on how "the conscious mind itself remains a *stranger* within that construct" (emphasis added). Perhaps there is a possibility of solving the HPE here, even if it would take on a strange form. Schrödinger clearly accepted that we are capable of motivated actions, and that this corresponds to physically measurable bodily changes in the material world that is described by the scientific world image. But if motivations are excluded from appearing as such in the domain of nature, then it follows that, within that domain, any such changes in bodily activity that are due to motivations would indeed appear as strange and unintelligible.

In other words, Schrödinger's exclusion principle essentially goes a step too far: what is excluded from the nature is not the *efficacy* of mind per se, but only the *understandability* of that efficacy for what it is, namely as changes originating from the mind. We can therefore propose a more moderate version of the exclusion principle based on the recognition that the principle of objectification forces the mind to make itself present in nature only in objectified form, which comes at the price that any motivated changes will thereby remain unintelligible from within that domain. On this view, there is no need to completely overcome the principle of exclusion; it is sufficient to allow that there are limits to the principle of understandability, which for Schrödinger had already been shown to be more malleable than classically expected before the quantum physics revolution. A moderate principle of exclusion therefore allows that the mind (motivations) can make a difference to matter (activity), thus opening a path toward solving the HPE, albeit only in a strange way: *when an organism's motivations make a difference to its physical activity, the principle of objectification entails that this difference falls outside of the applicability of the principle of understandability*. From the perspective of the physics of the organism, motivated changes will appear as correlated noise.

This crucial clarification opens a more nuanced view of the problem of mental causation. For example, it turns out that Schrödinger is only approximately correct when stating that the relation between mind and matter is such that mind "can neither act on it nor be acted on by any of its parts" (ibid., p. 119) – it would be more accurate to add the qualification: *such that it would also satisfy the principle of understandability*. This is a fundamentally different prospect for solving the HPE compared to ruling out all changes due to motivational involvement in our bodily activity altogether.

Similar considerations apply to the energetics of the organism. Here again we find that Schrödinger jumps from a strict interpretation of the exclusion principle to a questionable conclusion: "Sensations and thoughts do not belong to the 'world of energy', they cannot produce any change in this world of energy" (ibid., p. 127). Although it is correct that, like motivations more generally, sensations and thoughts do not belong in the domain of physics, and hence are not present *as such* in the world of energy, this does not entail that they cannot be associated with any *change* of energy in that domain. Rather, solving the HPE requires a more subtle claim to the effect there is indeed the possibility of a change in energy specifically due to activity that is motivated by the mind, but this change would go beyond the principle of understandability. One possibility is that part of the variability in the organism's fluctuating rate of entropy production could be attributed to motivation-correlated dissipation of free energy. We will unpack this possibility in more detail in the next two sections.

Schrödinger is not alone in grappling with the conundrum of the HPE in light of the quantum revolution. It is instructive to consider the closely aligned views of another Nobel laureate, Compton, who had won the 1922 Nobel Prize for Physics:

> A set of known physical conditions is not adequate to specify precisely what a forthcoming event will be. These conditions, insofar as they can be known, define instead a range of possible events from among which some particular event will occur. When one exercises freedom, by his act of choice he is himself adding a factor not supplied by the physical conditions and is thus himself determining what will occur. That he does so is known only to the person himself. From the outside one can see in his act only the working of physical law. It is the inner knowledge that he is in fact doing what he intends to do that tells the actor himself that he is free. (Compton, 1967)

A few differences between Compton's and Schrödinger's positions are worth noting. To begin with, whereas Schrödinger was still not entirely convinced that the probabilistic nature of quantum physics had demonstrated that there are limits to the principle of understandability, Compton takes indeterminacy as his starting point for considering questions regarding human freedom. Interestingly, Compton also goes a step further by allowing that choices made in the first-person perspective, by what Schrödinger called the "subject of cognizance", constitute an additional factor in its own right in the generation of action that goes beyond the physical description of the situation. But at the same time, and in apparent tension with the positing of this additional subjective factor, Compton also agrees with what Schrödinger called the principle of exclusion, given that Compton claims that "From the outside one can see in his act only the working of physical law". The tension is: how could motivated activity be fully accounted for by physical law, if the motivation of acting is also supposed to be "adding a factor not supplied by the physical conditions"? The solution may lie in the fact that the laws of nature Compton has in mind here are probabilistic rather than deterministic rules: "As far as physics is concerned, a person's actions which we think of as free would thus appear to occur simply according to the rules of chance."

However, this still brings us back to the HPE: on such a view, where an agent's activity is fully accounted for by the laws of physics, even if only statistically, there may not be sufficient room for motivations as such to make a difference to physical activity in their own right. Schrödinger is clear on this point: if the two principles of objectification and understandability are fully operational, the scientific world image simply cannot offer any solution to the HPE: for the alternative assumption – "the direct stepping in of free will to fill the gap of indeterminacy – does amount to an interference with the laws of nature, even in their form accepted in quantum theory. But at *that* price, of course, we can have everything. This is not a solution to the dilemma." (Schrödinger, 1996, p. 165). Of course, if we assume that anything goes, then the HPE would by definition no longer be a problem, but we would have lost all explanatory value of the solution at the same time.

In summary, recognition that the fundamental laws of nature are probabilistic is a step in the right direction to account for the efficacy of agency (Potter & Mitchell, 2022). Nevertheless, solving the dilemma of the HPE in a way that motivations as such can make a difference in their own right seems to require an even deeper makeover of the principle of understandability.

We can derive a few insights from the preceding discussion regarding what would be required to properly solve the HPE:

1) The principle of objectification remains strictly in place.
2) The principle of understandability can be relaxed in some cases, allowing for gaps in intelligible activity.
3) Yet it cannot be the case that anything goes; any revision of the principle of understandability to accommodate the efficacy of motivations must remain moderate. There will be limits to where and how motivations as such can make a difference to physical activity. For example:
    a. Only physical activity that is part of a living system can potentially be affected by motivations.
    b. Motivations cannot directly control how a gap in a living system's physical determination is filled.
    c. A living system's free energy places a limit on the extent to which motivations can make a difference.

In the following section we will present the irruption theory (Froese, 2023) as a development of the enactive approach to motivated activity that is broadly consistent with these criteria.

The thesis of irruption

We have arrived at an interpretation of Schrödinger's principles of the scientific method that makes room for a genuine solution to the HPE. To recap: when an organism's motivations make a difference to its activity, this will not show up as motivated as such (due to the *principle of objectification*), yet it will still be associated with a physically measurable change (because the *exclusion principle* is sufficiently relaxed, thereby softening the HPE), albeit that change will remain in principle unintelligible from the domain of physics (due to a simultaneous relaxation of the *principle of understandability*). This reformulation of the principle of exclusion as pertaining to intelligibility, rather than efficacy per se, has the benefit that it opens the possibility of a scientifically workable solution to the HPE. What it calls for is a new interpretation of noise generation that is correlated with the motivated activity of living systems.

This call aligns with a recent proposal to advance the enactive approach to motivated activity by operationalizing an agent's motivational involvement in the generation of its behavior in terms of increased entropy production (Barrett, 2020a, 2020b; Swenson, 2020). Froese (2023) responded with an analysis of the motivational involvement as associated with an increase in the underdetermination of that activity's constitutive material processes. The burst of underdetermined activity is called an "irruption" and it is argued to be approximately measurable by means of the resulting increase in fluctuations, specifically with information-theoretic measures of entropy.

This is not the place to assess the arguments and evidence that support the thesis of irruption in detail; the interested reader is referred to the exposition by Froese (2023). In the context of the current discussion, it should suffice to conceive of irruption as a motivated change in an organism's physical activity that is thereby not subject to the principle of understandability. Still, it is worth highlighting that solving the HPE in terms of the appearance of irruptions in the physical domain is consistent with growing experimental evidence. For example, increases in information-theoretic entropy of neural activity has already been consistently linked with increased levels of cognitive and behavioral activity (Lynn, Cornblath, Papadopoulos, Bertolero, & Bassett, 2021), and with increased levels of subjectively reported awareness (Schartner, Carhart-Harris, Barrett, Seth, & Muthukumaraswamy, 2017). In addition, there are compelling reasons to re-interpret the "readiness potential" associated with the self-initiation of movement in terms of the accumulation of stochastic fluctuations in neural activity (Schurger, Hu, Pak, & Roskies, 2021). In theoretical terms, the thesis of irruption is also a systematic response to Froese and Taguchi's (2019) critical assessment that enactive AI, artificial life, as well as the enactive approach more generally (Froese, 2018), need to make room for a role of indeterminacy in order to better account for the possibility of motivated behavior.

The immediate worry that the relinquishing of determinism only entails its opposite extreme, that is, completely random behavior, is well known from the libertarian philosophy of free will (Tewes, 2017). However, the field of artificial life provides a rich source of theory and models that demonstrate how arbitrary changes can actually operate in the service of adaptive behavior (e.g. Watson, Mills, & Buckley, 2011). The enactive approach has already long been inspired by relevant pioneering work in the cybernetics era, especially by Ashby's (1960) stochastic implementation of ultrastability (Di Paolo, 2010). At the core of ultrastability lies the concept of a sudden "break" in the system's organization, which refers to an arbitrary step-change in organization following an environmental perturbation that has pushed an essential variable beyond its boundary of viability. But these breaks depend on the environment, and so do not readily account for the more active, homeodynamic features of living systems (Ikegami & Suzuki, 2008). In this sense, the thesis of irruption can be interpreted as an extension of Ashby's work in the direction of a kind of self-breaking system.

More recently, a line of research in artificial life initiated by Watson and colleagues (2011) on modeling complex adaptive systems in terms of arbitrary state resets and unsupervised learning has been re-interpreted by Froese and colleagues in terms of a Self-Optimization Model of the enactive conception of life (Froese, Weber, Shpurov, & Ikegami, 2023). Briefly, the enactive conception of life proposed by Di Paolo et al. (2017) holds that a living system must adaptively satisfy the primordial tension between (1) being sufficiently open to the environment so as to maintain free energy with which to do work, and (2) sufficiently closed to the environment to maintain its structural integrity. These two essential requirements cannot be satisfied simultaneously, and their respective constraints must therefore be coordinated across time. Modeling the system's transient openness as the breaking of detailed balance of the closed system's equilibrium, which is implemented by means of an arbitrary reset of system state, plus adding plasticity in the form of a slow sedimentation of previously visited equilibrium states, enables the system to self-optimize its organization such that it better satisfies its constraints (Froese et al., 2023). This situation is illustrated in Figure 1 below.

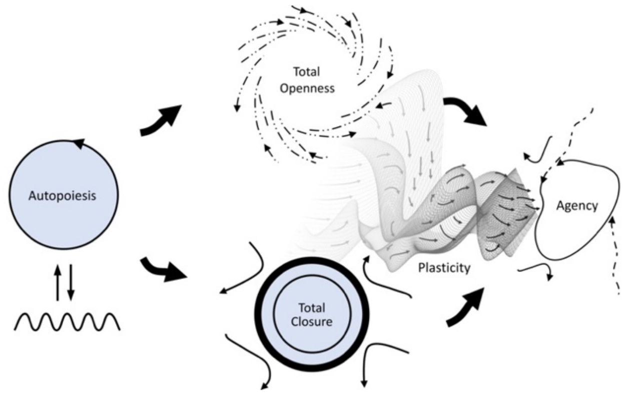

Figure 1. The enactive conception of life by Di Paolo et al. (2017) highlights that there is a primordial tension at the core of autopoiesis: a far-from-equilibrium system that self-produces its own identity must open itself to the environment because it needs energy, and close itself off from the environment to maintain its structural integrity. If it tends too much toward satisfying either requirement it will dissolve or starve, respectively. A living system must therefore maintain a precarious balance by adaptively coordinating the partial satisfaction of the two opposing needs over time. The Self-Optimization Model by Froese et al. (2023) illustrates how this kind of agency can come about, namely by introducing plasticity in the form of associative learning. Figure originally from Froese et al. (2023), based on Di Paolo et al. (2017).

This adaptive role of the repeated breaking of the system's detailed balance makes a good fit with the thesis of irruption, which includes two complementary theses: the thesis of scaling and the thesis of attunement. Together they account for the transformation of irruptions happening at the smallest scale of description into behavior that is in alignment with the agent's motivations. This involves scaling up irruption to the system level, for example due to chaos or self-organized criticality, in the context of a living system that is poised to switch from one behavioral attractor to another (Froese, 2023). Future research could therefore unpack the possibility that irruptions could play a similar "reset" function in the Self-Optimization model to that of the system transiently opening to the external environment, which similarly pushes it out of its closed equilibrium state. The overarching key insight is that such "breaks" in the system's organization or state are not detrimental to, but rather facilitate, the realization of adaptive behavior.

All three theses regarding how to solve the HPE – irruption, scaling, and attunement – therefore make suitable targets for theoretical development by means of artificial life modeling. Nevertheless, before irruptions can be modeled in a less abstract form, fundamental questions must be addressed about how to conceive of the irruption of an agent's motivations into its behavior's material basis. This is not an easy task, given that the physics of life and the thermodynamics of information are still in early stages of development (Davies, 2019). However, given that the special theme of this artificial life conference is precisely to revisit and update the classic idea of the "ghost in the machine", it is appropriate to offer some considerations that may take us a small step in that direction.

Toward a physics of irruption

We can start to develop the thermodynamic implications of irruptions based on Schrödinger's work, especially his 1943 lectures *What is life? The physical aspects of the living cell*. An important theme of those lectures is that organisms remain ordered as long as they work to maintain themselves far from thermodynamic equilibrium, because falling into equilibrium with their environment is equivalent with dissolution and death. Organisms manage to accomplish this because of metabolism: they make available free energy inside their body by consuming free energy from the environment, and they also return disorder to the environment which increases its entropy. This enables them to dissipate free energy in useful ways, for example by doing work that maintains their ordered, far-from-equilibrium state, leading to a work-constraint cycle (Kauffman, 2000).

How would irruptions that are associated with an organism's motivated activity make a difference to this thermodynamic situation? As discussed above, given Schrödinger's inviolable principle of objectification, any changes in physical activity that are due to motivational involvement as such could only manifest as breaking the principle of understandability, that is, by showing up as disorder or noise that is correlated with the motivated activity. In other words, irruptions would be an extra factor of variability of entropy production.

This variable noise term could be beneficial, as noise plays a variety of functional roles in organisms (Ilan, 2023). To give just one example, it may make it more probable that the system surpasses energy thresholds standing between one metastable state and another, thereby resulting in irreversible changes. Such irreversibility is of interest, because under a broad range of conditions heat dissipation can facilitate the emergence of adaptive behavior in self-assembling systems (England, 2015).

In other words, we can think of irruptions as introducing a bias to the state-to-state transitions of a living system, similar to certain treatments of Maxwell's demon (Parrondo, Horowitz, & Sagawa, 2015), yet where the demon's capacity to create ordered transitions by imposing constraints on state dynamics is replaced by the weaker ability to indirectly facilitate state transitions by imposing bursts of disorder. If this is on the right track, it would entail a revision of the thermodynamics of the living compared to what had been proposed by Schrödinger, which has so far remained largely unchallenged in its essential outlook. There would be something that differentiates entropy production of living from non-living dissipative structures: the irruption of disorder correlated with motivated activity.

Again, the source of this disorder is due to the objectification of the differences that subjective factors make to the organism's behavior. This objectification amounts to a simplification of reality, as Schrödinger had recognized (1992, p. 118); hence, the source of this disorder is necessarily unintelligible from the domain of physics. This implication of a relaxed principle of understandability will not be to everyone's liking. An important question for future work is therefore to assess just how much is at stake: is there is a way of conceiving of irruptions such that they are not in tension with probabilistic physical laws? Can we formulate the far-from-equilibrium thermodynamics of living systems such that occasional arbitrary departures from average thermodynamic tendencies due to irruptions are permitted?

If this cannot be done, this proposed solution to the HPE would be a hard sell for many. For example, Rovelli (2021) eexcludes this kind of possibility from his considerations of the physics of agency: “The existence of agency that does not to satisfy neither deterministic nor probabilistic physical laws (*supernatural*) would contradict our current understanding of our world I see no interest in considering it.” However, this rejection is too strong. The ambition should indeed be to satisfy all physical laws; yet making room for a conception of agency that solves the HPE could still require a moderate relaxation of the principle of understandability. This relaxation needs to go hand in hand with a broader conception of nature to include the “world beyond physics”, especially life and mind (Kauffman, 2019). In this respect the proposal of irruption theory is in line with ongoing developments of a relaxed naturalism in enactive cognitive science (Hutto & Satne, 2018a). Similarly, we have seen that, in contrast to Rovelli, Schrödinger and Compton had struggled with the recognition that there is more to our agency than what the principle of objectification allows to enter into the scientific construct of the domain of nature. And perhaps it was this awareness of the simplifying filter on nature imposed by the principle of objectification that made it easier for them to consider the very possibility that there may be limits to the principle of understandability in physics. The thesis of irruption is well-placed to become the theoretical starting point from which to develop these ideas.

As a starting point, Swenson (2020) has recently defended the thermodynamic foundations of ecological psychology by appealing to an agent’s capacity for arbitrary change as a way of distinguishing specifically cognitive systems from the more general class of self-producing autocatakinetic systems: “The hallmark of cognitive systems is that they can constitute their autocatakinetics in ways that are arbitrary to and often antithetical to local force-fields or potentials.” Yet the details of this capacity remain unclear. How does the system sidestep the constraints imposed by its energetics? While agreeing with Swenson’s distinction, Barrett (2020a, p. 117) therefore notes that “something more needs to be said about the organization of living stems that allows for this arbitrariness with respect to local potentials.” This question about the source of arbitrariness is interesting from the perspective of irruption theory, because the disorder it posits to be associated with motivated activity may be specifically suited as an answer. This characterization of arbitrariness will require more careful unpacking by future work, but we can at least point in the general direction.

In the Boltzmann machine, a spontaneous switch between metastable attractors is facilitated by higher system temperature because this makes its state changes more erratic; hence, it will become more likely that the system makes spontaneous jumps up its energy gradient, compared to its natural tendency of following the gradient downwards toward thermodynamic equilibrium. Importantly, the surface of the energy landscape that defines the local gradient remains unchanged.

According to irruption theory, however, the thermodynamic situation of a living system’s motivated activity may be subtly different. Similar to increasing temperature, motivated activity is associated with more disorder of state changes. However, the source of this disorder is more fundamental: the state changes manifest as disordered because they are additionally affected by agential factors that go beyond factors describable purely in terms of the physical constraints of the system. In other words, irruptions have the effect of disordering state changes in a way that is arbitrary with respect to the system’s energy gradients – it is as if the whole energy landscape is temporarily be modified to enable a random walk in state space. If this idea can be more formally substantiated, it would mean that irruptions could provide living systems with additional degrees of freedom: the arbitrariness resulting from irruptions implies that, overall, the living system’s changes of state would be less constrained by its energy barriers, compared to mere increases in temperature.

Yet one key thermodynamic constraint remains the same: just like normal thermodynamic work done to produce order in the living system, which produces heat as a side product, the arbitrary disorder of irruptions adds pressure on the organism to renew its energy budget. Simply put, the thesis of irruption points to the novel hypothesis that *an organism’s motivational involvement in the generation of its behavior is associated with its own specific cost of thermodynamic free energy*. Mental work is similar to mechanical work in this respect.

This hypothesis fits nicely with recent work on the feeling of effort, which converges on the claim that the feeling of effort is one and the same for both mental and bodily actions, and that it is associated with subjective cost of inhibiting an automatic tendency to stop the action (Bermúdez, in press). This links the feeling of effort with the system’s thermodynamic energy cost of deviating from the principle of least action.

To some extent this is an intuitive correspondence, which fits with the physics of life more generally. Metabolic flux has the right kind of characteristics to be associated with the subjective dimension of life and mind (Lane, 2022). To take a well-known example, it is common practice to track cognitive activity in terms of brain energy requirements. This is what motivated the development of the standard technique of Blood Oxygenation Level Dependent (BOLD) brain imaging. However, irruption theory suggests that there would be value in broadening the current focus to also look for unexpected drains on energy. We can assume that such empirical findings, like anything else that would require relaxation of the principle of understandability, will be seriously underreported in the literature. Still, there are curious high-profile examples in the literature that suggest this is worthy of more systematic investigation.

First, it is a widely remarked puzzle just how much energy the brain uses compared to the rest of the body, even when it is supposedly in a resting state (Raichle, 2006). In fact, even task-related BOLD signals are not restricted to energy expenditure related to neural spiking; there are blood vessels that become robustly activated in response to stimuli, even if those stimuli evoked little to no neural spiking activity in the surrounding tissue (O’Herron et al., 2016). Strikingly, even neurons that are not electrically active must deal with a major metabolic burden, as the maintenance of synaptic vesicle pools in nerve terminals accounts for up to 44% of resting synaptic energy consumption (Pulido & Ryan, 2021). Pulido and Ryan offer an intriguing interpretation of this metabolic burden: “Faster transport is generally achieved by lowering the energy barrier for the key structural transition, which, in turn, implies a higher probability of occurring from thermal fluctuations.” In the light of irruption theory, we may speculate whether neurons’ sensitivity to tiny noise fluctuations could be a feature rather than a bug.

So far, we have been considering implications of irruption theory for the thermodynamics of an individual living system. However, it may also be possible to detect free energy loss due to irruptions when it is measured as accumulated over a much larger scale, for example at the level of whole ecosystems and especially of the whole biosphere at a planetary scale. Fittingly, ecological psychology has long been developing links between the apparent tendency of organisms to approximate conditions of maximum entropy production with life's capacity for open-ended evolution of complexity (Swenson & Turvey, 1991). More recently, there are attempts at explaining evolution of complexity by linking thermodynamic entropy production to the generation of information-theoretic entropy associated with the opening of new cognitive state spaces (Jeffery, Pollack, & Rovelli, 2019). In this regard, irruption theory could help to address how the emergence of such informational state spaces for cognitive systems could make a difference that matters in its own right to the thermodynamics of life (Davies, 2019).

Concluding remarks

Schrödinger rejected the standard assumption that our mind is located inside the head of our physical bodies – internalism – because it forms part of a long chain of faulty conclusions about the relationship between mind and matter. This rejection of an internalist and representationalist view of the mind is shared by the enactive approach. However, the allure of internalism is sufficiently powerful that this assumption continues to form the default starting point for most of cognitive science today, as it seems to do for most of physics. Consider the following start to a book that aims to rebuild physics from basic principles:

> A hard lesson to learn is that our sensations are partly caused by reality, but are fully constructed by our brains to present the world to us in just the form we need to make our way in nature. Beyond those sensations, nature hovers, fundamentally mysterious and just at the edge of what we can know. (Smolin, 2019, p. xiii)

Although the enactive approach would agree that the domain of nature transcends our mind's grasp, it rejects this neat causal separation of internal and external, and of conscious mind and material body. Instead, it is developing a more participatory realism that encompasses inner and outer, and involves both experience and world (Froese, 2022). Schrödinger, on the other hand, ended up defending the opposite extreme of internalist representationalism, namely a kind of holistic idealism:

> The reason why our sentient, percipient and thinking ego is met nowhere within our scientific world picture can easily be indicated in seven words: because it is itself that world picture. It is identical with the whole and therefore cannot be contained in it as a part of it. (Schrödinger, 1992, p. 128)

In the different lecture series Schrödinger is consistently attracted to mystical themes that identify the material world with, or at least place it within, consciousness. This fascination with certain strands of the perennial philosophy perhaps helps to explain his reluctance to accept limits to the principle of understandability, which would indicate that, to some extent, nature transcends consciousness after all.

However, neither the materialist assumption of complete separation nor the idealist assumption of complete identity does proper justice to the complex interdependence between mind and matter, and the resulting cross-border effects (Wagemann, 2011). The enactive approach has therefore been developing an alternative framing. Even before its inauguration in the 90s, Varela (1976, 1984) had been inspired by the Zen principle of nonduality, which holds that the relation between mind and body is such that they are "not one, not two". That is, mind and body do participate in each other's specific domains, yet they do so without losing their own unique characteristics. This nondual approach to conceptualizing the intersection of these two domains of reality, that is, in a manner that is mutually respectful of their ontological characteristics, is reminiscent of an engaged epistemology of "letting be" (De Jaegher, 2021). It calls for a further development of relaxed naturalism (Hutto, 2023): we can only properly grasp mind-body interaction in our understanding by at the same time partially loosening the grip of our strict naturalistic understanding of that relation.

This scientific "letting be" requires a willingness to open up traditional disciplinary boundaries that are heavily defended, and to relax the principle of understandability. Yet there is a notable payoff for revising the basic principles of the scientific method so that it can overcome the HPE of motivated activity. The upshot is that, in contrast to Schrödinger's strict exclusion principle, we are not impotent with respect to the material world that is disclosed by science. In line with more agency-centered takes on quantum physics (Barad, 2007; C. A. Fuchs, 2017), we meet the material world halfway – we can make a difference, albeit not exactly to our choosing.

This middle way between freedom and determinism may leave some staunch libertarians unsatisfied. However, it has the benefit of offering a scientifically workable research program. If motivated activity is correlated with irruptions, then this leads to the prediction of a physical trace that would otherwise not be there. Specifically, this trace should be measurable as an increase in entropy production.

Finally, the thesis of irruption raises the intriguing prospect for a fundamental reassessment of the methods and ambitions of artificial life. For instance, if it is accepted that neither the complete separation nor the complete identity of mind and body can be upheld, what does this mean for attempts at synthetically realizing the mind purely in terms of body? Simulation models then start to look even more limited in scope in terms of what insights they can provide. But even "wet" approaches that do not struggle with the limitations of virtuality need a rethink to have better chances of giving rise to motivated activity. Are these chemical systems sufficiently open and complex such that irruptions due to regulation of viability constraints could occur, and even become the driving force of the development of adaptive behavior, akin to the use of arbitrary state resets in the Self-Optimization model (Froese et al., 2023)?

It is also worth reflecting about what an alternative synthetic approach could look like, especially if it needs to be based more on an engaged epistemology. In this regard, a recent proposal to employ human participants for motivated interaction with android systems to facilitate their own personogenesis via mind contagion (Masumori, Maruyama, & Ikegami, 2021) might offer clues to the future of artificial life.

References

Agmon, E., Gates, A. J., & Beer, R. D. (2016). The structure of ontogenies in a model protocell. *Artificial Life, 22*, 499-517.

Ashby, W. R. (1960). *Design for a Brain: The Origin of Adaptive Behaviour* (Second ed.). London, UK: Chapman & Hall.

Barad, K. (2007). *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Durham: Duke University Press.

Barandiaran, X., & Egbert, M. D. (2014). Norm-establishing and norm-following in autonomous agency. *Artificial Life, 20*(1), 5-28.

Barbaras, R. (2010). Life and exteriority: The problem of metabolism. In J. Stewart, O. Gapenne, & E. A. Di Paolo (Eds.), *Enaction: Toward a New Paradigm for Cognitive Science* (pp. 89-122). Cambridge, MA: The MIT Press.

Barrett, N. (2017). The normative turn in enactive theory: An examination of its roots and implications. *Topoi, 36*, 431-443.

Barrett, N. (2020a). Extremal properties and self-preserving behavior. *Adaptive Behavior, 28*(2), 113-118.

Barrett, N. (2020b). On the nature and origins of cognition as a form of motivated activity. *Adaptive Behavior, 28*(2), 89-103.

Beer, R. D. (1997). The dynamics of adaptive behavior: A research program. *Robotics and Autonomous Systems, 20*, 257-289.

Bermúdez, J. P. (in press). What is the feeling of effort about? *Australasian Journal of Philosophy*.

Bourgine, P., & Varela, F. J. (1992). Introduction: Towards a practice of autonomous systems. In F. J. Varela & P. Bourgine (Eds.), *Toward a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life* (pp. xi-xvii). Cambridge, MA: The MIT Press.

Cleeremans, A., & Tallon-Baudry, C. (2022). Consciousness matters: Phenomenal experience has functional value. *Neuroscience of Consciousness, 2022*(1), niac007. doi:10.1093/nc/niac007

Compton, A. H. (1967). *The Cosmos of Arthur Holly Compton*. New York: Knopf.

Conway, J. H., & Kochen, S. (2009). The strong free will theorem. *Notices of the American Mathematical Society, 56*(2), 226-232.

Davies, P. C. W. (2019). *The Demon in the Machine: How Hidden Webs of Information Are Solving the Mystery of Life*. Chicago, IL: Chicago University Press.

De Jaegher, H. (2021). Loving and knowing: Reflections for an engaged epistemology. *Phenomenology and the Cognitive Sciences, 20*, 847-870.

Di Paolo, E. A. (2003). Organismically-inspired robotics: Homeostatic adaptation and teleology beyond the closed sensorimotor loop. In K. Murase & T. Asakura (Eds.), *Dynamical Systems Approach to Embodiment and Sociality* (pp. 19-42). Adelaide, Australia: Advanced Knowledge International.

Di Paolo, E. A. (2010). Robotics inspired in the organism. *Intellectica, 1-2*(53-54), 129-162.

Di Paolo, E. A., Buhrmann, T., & Barandiaran, X. (2017). *Sensorimotor Life: An Enactive Proposal*. Oxford, UK: Oxford University Press.

Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Philosophical Psychology, 20*(2), 247-268.

Egbert, M. D., & Barandiaran, X. (2014). Modeling habits as self-sustaining patterns of sensorimotor behavior. *Frontiers in Human Neuroscience, 8*(590). doi:10.3389/fnhum.2014.00590

England, J. L. (2015). Dissipative adaptation in driven self-assembly. *Nature Nanotechnology, 10*. doi:10.1038/NNANO.2015.250

Froese, T. (2018). Ecology of the Brain: The Phenomenology and Biology of the Embodied Mind. *Frontiers in Psychology, 9*, 2174. doi:10.3389/fpsyg.2018.02174

Froese, T. (2022). Scientific observation is socio-materially augmented perception: Toward a participatory realism. *Philosophies, 7*(2), 37. doi:10.3390/philosophies7020037

Froese, T. (2023). Irruption Theory: A novel conceptualization of the enactive account of motivated activity. *Entropy, 25*(5), 748. doi:10.3390/e25050748

Froese, T., & Taguchi, S. (2019). The problem of meaning in AI and robotics: Still with us after all these years. *Philosophies, 4*, 14. doi:10.3390/philosophies4020014

Froese, T., Weber, N., Shpurov, I., & Ikegami, T. (2023). From autopoiesis to self-optimization: Toward an enactive model of biological regulation. *bioRxiv, 2023.02.05.527213*. doi:10.1101/2023.02.05.527213

Froese, T., & Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence, 173*(3-4), 366-500.

Fuchs, C. A. (2017). On participatory realism. In I. T. Durham & D. Rickles (Eds.), *Information and Interaction* (pp. 113-134). Switzerland: Springer.

Fuchs, T. (2021). *In Defense of the Human Being: Foundational Questions of an Embodied Anthropology*: Oxford University Press.

Gallagher, S. (2017). *Enactivist Interventions: Rethinking the Mind*. Oxford, UK: Oxford University Press.

Gallagher, S. (2018). Rethinking nature: Phenomenology and a non-reductionist cognitive science. *Australasian Philosophical Review, 2*(2), 125-137.

Hanna, R., & Maiese, M. (2009). *Embodied Minds in Action*. Oxford: Oxford University Press.

Hanna, R., & Thompson, E. (2003). The mind-body-body problem. *Theoria et Historia Scientarum, 7*(1), 23-42.

Hutto, D. D. (2023). A positively relaxed take on naturalism: Reasons to be relaxed but not too liberal. *Topoi*. doi:10.1007/s11245-023-09892-8

Hutto, D. D., & Myin, E. (2013). *Radicalizing Enactivism: Basic Minds without Content*. Cambridge, MA: The MIT Press.

Hutto, D. D., & Myin, E. (2017). *Evolving Enactivism: Basic Minds Meet Content*. Cambridge, MA: MIT Press.

Hutto, D. D., & Satne, G. (2018a). Naturalism in the Goldilocks Zone: Wittgenstein's delicate balancing act. In K. M. Cahill & T. Raleigh (Eds.), *Wittgenstein and Naturalism* (pp. 56-76). New York, NY: Routledge.

Hutto, D. D., & Satne, G. (2018b). Wittgenstein's inspiring view of nature: On connecting philosophy and science aright. *Philosophical Investigations, 41*(2), 141-160.

Ikegami, T., & Suzuki, K. (2008). From homeostatic to homeodynamic self. *BioSystems, 91*(2), 388-400.

Ilan, Y. (2023). Constrained disorder principle-based variability is fundamental for biological processes: Beyond biological relativity and physiological regulatory networks. *Progress in Biophysics and Molecular Biology, 180-181*, 37-48. doi:https://doi.org/10.1016/j.pbiomolbio.2023.04.003

Jeffery, K. J., Pollack, R., & Rovelli, C. (2019). On the statistical mechanics of life: Schrödinger revisited. *Entropy, 21*, 1211. doi:10.3390/e21121211

Kauffman, S. A. (2000). *Investigations*. New York, NY: Oxford University Press.

Kauffman, S. A. (2019). *A World Beyond Physics: The Emergence and Evolution of Life*. New York, NY: Oxford University Press.

Lane, N. (2022). *Transformer: The Deep Chemistry of Life and Death*. New York, NY: W. W. Norton & Company, Inc.

Lynn, C. W., Cornblath, E. J., Papadopoulos, L., Bertolero, M. A., & Bassett, D. S. (2021). Broken detailed balance and entropy production in the human brain. *Proceedings of the National Academy of Sciences, 118*(47), e2109889118. doi:10.1073/pnas.2109889118

Masumori, A., Maruyama, N., & Ikegami, T. (2021). Personogenesis through imitating human behavior in a humanoid robot "Alter3". *Frontiers in Robotics and AI, 7*, 532375. doi:10.3389/frobt.2020.532375

Myin, E., & Zahnoun, F. (2018). Reincarnating the identity theory. *Frontiers in Psychology, 9*(2044). doi:10.3389/fpsyg.2018.02044

O'Herron, P., Chhatbar, P. Y., Levy, M., Shen, Z., Schramm, A. E., Lu, Z., & Kara, P. (2016). Neural correlates of single-vessel haemodynamic responses in vivo. *Nature, 534*(7607), 378-382. doi:10.1038/nature17965

Parrondo, J. M. R., Horowitz, J. M., & Sagawa, T. (2015). Thermodynamics of information. *Nature Physics, 11*(2), 131-139. doi:10.1038/nphys3230

Potter, H. D., & Mitchell, K. J. (2022). Naturalising agent causation. *Entropy, 24*(4). doi:10.3390/e24040472

Pulido, C., & Ryan, T. A. (2021). Synaptic vesicle pools are a major hidden resting metabolic burden of nerve terminals. *Science Advances, 7*(49), eabi9027. doi:doi:10.1126/sciadv.abi9027

Raichle, M. E. (2006). The brain's dark energy. *Science, 314*(5803), 1249-1250. doi:doi:10.1126/science.1134405

Ramírez-Vizcaya, S., & Froese, T. (2020). Agents of habit: Refining the artificial life route to artificial intelligence. In J. Bongard, J. Lovato, L. Hebert-Dufrésne, R. Dasari, & L. Soros (Eds.), *ALIFE 2020: The 2020 Conference on Artificial Life* (pp. 78-86). Cambridge, MA: MIT Press.

Rovelli, C. (2021). Agency in physics. In C. Calosi, P. Graziani, D. Pietrini, & G. Tarozzi (Eds.), *Experience, Abstraction and the Scientific Image of the World* (pp. 25-40): FrancoAngeli.

Schartner, M. M., Carhart-Harris, R. L., Barrett, A. B., Seth, A. K., & Muthukumaraswamy, S. D. (2017). Increased spontaneous MEG signal diversity for psychoactive doses of ketamine, LSD and psilocybin. *Scientific Reports, 7*(46421). doi:10.1038/srep46421

Schrödinger, E. (1992). *What Is Life? with Mind And Matter and Autobiographical Sketches*. Cambridge, UK: Cambridge University Press.

Schrödinger, E. (1996). *Nature and the Greeks and Science and Humanism*: Cambridge University Press.

Schurger, A., Hu, P. B., Pak, J., & Roskies, A. L. (2021). What Is the Readiness Potential? *Trends in Cognitive Sciences, 25*(7), 558-570.

Smolin, L. (2019). *Einstein's Unfinished Revolution: The Search for What Lies Beyond the Quantum*. New York: Penguin Press.

Swenson, R. (2020). The fourth law of thermodynamics (LMEP) and cognition from first principles: Commentary on Barrett's "On the nature and origins of cognition as a form of motivated activity". *Adaptive Behavior, 28*(2), 105-107.

Swenson, R., & Turvey, M. T. (1991). Thermodynamic reasons for perception-action cycles. *Ecological Psychology, 3*(4), 317-348.

Tewes, C. (2017). *Libertarismus, Willensfreiheit und Verursachung*. Frankfurt am Main, Germany: Vittorio Klostermann.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Harvard University Press.

van Buuren, J. (2018). *Body and Reality: An Examination of the Relationships between the Body Proper, Physical Reality, and the Phenomenal World starting from Plessner and Merleau-Ponty*. Bielefeld: transcript.

Varela, F. J. (1976). Not one, not two. *The Co-Evolution Quarterly, 12*, 62-67.

Varela, F. J. (1984). The creative circle: Sketches on the natural history of circularity. In P. Watzlawick (Ed.), *The Invented Reality* (pp. 309-324). New York, NY: W. W. Norton & Company, Inc.

Varela, F. J. (1991). Organism: A meshwork of selfless selves. In A. I. Tauber (Ed.), *Organism and the Origins of Self* (pp. 79-107). Dordrecht, Netherlands: Kluwer Academic Publishers.

Varela, F. J. (1995). The re-enchantment of the concrete: Some biological ingredients for a nouvelle cognitive science. In L. Steels & R. Brooks (Eds.), *The Artificial Life Route to Artificial Intelligence* (pp. 11-22). Hove, UK: Lawrence Erlbaum Associates.

Varela, F. J. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies, 3*(4), 330-349.

Varela, F. J., Maturana, H. R., & Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. *BioSystems, 5*, 187-196.

Varela, F. J., & Thompson, E. (2003). Neural synchrony and the unity of mind: A neurophenomenological perspective. In A. Cleeremans (Ed.), *The Unity of Consciousness: Binding, Integration, and Dissociation* (pp. 266-287). New York, NY: Oxford University Press.

Varela, F. J., Thompson, E., & Rosch, E. (2017). *The Embodied Mind: Cognitive Science and Human Experience* (Revised ed.). Cambridge, MA: MIT Press.

Villalobos, M. (2013). Enactive cognitive science: Revisionism or revolution? *Adaptive Behavior, 21*(3), 159-167.

Villalobos, M., & Ward, D. (2016). Lived experience and cognitive science: Reappraising enactivism's Jonasian turn. *Constructivist Foundations, 11*(2), 204-212.

Wagemann, J. (2011). The structure-phenomenological concept of brain-consciousness correlation. *Mind & Matter, 9*(2), 185-204.

Watson, R. A., Mills, R., & Buckley, C. L. (2011). Global adaptation in networks of selfish components: Emergent associative memory at the system scale. *Artificial Life, 17*(3), 147-166.

Zahavi, D. (2017). *Husserl's Legacy: Phenomenology, Metaphysics, and Transcendental Philosophy*. Oxford, UK: Oxford University Press.