**Making mind matter with irruption theory: Bridging end-directedness and entropy production by satisfying the participation criterion**

Tom Froese 1 \*, Georgii Karelin 1, and Takashi Ikegami 2, 3

\* Corresponding author’s email: tom.froese@oist.jp

1 Embodied Cognitive Science Unit, Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Okinawa 904-0495, Japan

2 Ikegami Lab, Graduate School of Arts and Sciences, University of Tokyo, 3-8-1 Komaba, Tokyo 153-8902, Japan

3 Theoretical Sciences Visiting Program (TSVP), Okinawa Institute of Science and Technology Graduate University, Onna, 904-0495, Japan

**Abstract**

Biological processes are end-directed, that is, teleological. Explaining the physical efficacy of end-directedness continues to be a profound challenge for theoretical biology, especially given its unavoidable implications for our own self-understanding. For a comprehensive theory of life, it is pivotal to bridge our human-centric view of end-directedness, which the social sciences and humanities consider intrinsic to our actions, with the natural sciences’ view of actions’ in purely physiological terms, especially in terms of thermodynamic tendencies. A comprehensive theory should therefore provide an end-involving account, which illuminates how both physiology and teleology distinctly contribute to behavior generation. Here we introduce the “Participation Criterion”: *End-involvement in a bodily process entails that, in principle, it is distinguishable from one without end-involvement, specifically in terms of physiologically unpredictable changes in unexplainable variability*. To exemplify the difficulty of satisfying this criterion, we critically analyze two theories on the thermodynamic basis of end-directedness. We then propose that “Irruption Theory” points to a way forward because it predicts that bodily processes have an end-involvement-dependent increase in their entropy rate. This is consistent with evidence of an association between conscious intention and neural fluctuations, is open to further experimental verification, and provides a novel perspective on the role of thermodynamic entropy production in the organism.

**Keywords:** maximum entropy production principle; dissipative structures; autogen; autocatakinetics; ecological psychology; agency

1. **Introduction**

One of the most active frontiers of scientific knowledge is the question of how conscious agents like ourselves fit into the natural order described by the physical and life sciences (e.g. Azarian, 2022; Mitchell, 2023; Musser, 2023; Sapolsky, 2023). Questions regarding the possibility of teleology, denoting the inherent goal- or end-directedness of biological processes, become even more intricate when considering the intentional actions of humans (Noble & Ellis, 2022). This notion—that the mental capacity for aiming at future potentials can influence present behavioral realities—has historically challenged scientific thought. Moreover, doubts over the causal efficacy of end-directedness have engendered a rift between the natural and social sciences. While the natural sciences often dismiss “ends” as ultimately inconsequential for the most fundamental scientific explanations of reality (Carroll, 2022), the humanities and social sciences uphold them as essential for the very possibility of fundamental science (Kim, 2005). Both stances, in their extremes of elimination versus presupposition, make it unintelligible how end-involvement might manifest in the intermediate scales of biology or cognitive science.

However, throughout modern history, some have cautioned against such a stark division of the natural and social sciences, hinting at deeper connections between mind and matter. Nearly a century ago, the physical chemist Watson (1930, p. 222) aptly captured this sentiment, stating: “Thought interferes with the probability of events, and, in the long run, therefore, with entropy.” However, Watson seemed to have envisioned establishing this connection between thought and entropy by abstracting an information-theoretic concept of entropy away from concrete thermodynamic considerations related to changes in the body’s available energy to do work. This information-theoretic ‘disembodiment’ of mind was also a foundational principle for the subsequent cybernetics era, which deployed it in support of arguing for a systems-theoretic equivalence between animals and machines (Wiener, 1961). In hindsight this abstraction, although useful for stimulating more interdisciplinary approaches to life and mind, may turn out to have ultimately been a distraction from potentially deeper connections between mind and matter at the level of thermodynamics. For example, over recent decades, there has been a rise in efforts to elucidate the end-involving activity of living beings in specifically thermodynamic terms (e.g., Deacon, 2012; Swenson & Turvey, 1991; Tschacher & Haken, 2007).

Yet, like the causal exclusion principle haunting nonreductive physicalism more generally (Kim, 2005), these thermodynamic accounts of end-directedness have a notable shortcoming: the assumed presence of end-involvement is, empirically, indistinct from its assumed absence[[1]](#footnote-1). If we accept the premise that ends supervene on bodily processes in a world characterized by physical causal completeness at the scale of physiology or smaller, appeals to end-involvement can at best be a heuristic or epistemic tool, arguably redundant in essence. As such, the challenge is to find a formulation of end-involvement that does justice to the most compelling insights that both the natural and social sciences have to offer, and in such a way that neither rules out what is specific to the domain of the other.

Arguably, this ontological complexity is not accidental, but an essential and irreducible aspect of our ambiguous being in the world, as long noted by phenomenologically minded thinkers: “our bodies are both subjects open to the things surrounding us, and themselves such things” (van Buuren, 2018, p. 34). Accordingly, our task is to describe the efficacy of teleology as a special relationship between two intrinsically unlike domains of phenomena or entities, but in such a way that neither can be excluded, and hence offering a genuine possibility of irreducible cross-domain effects (Nicolescu, 2012; Wagemann, 2011).

We can reformulate this reciprocal respect for domain-specific characteristics into a criterion to assess how successful an account of behavior generation is in securing end-involvement. For purposes of illustration, let us work with an example proposed by Moore in the context of the debate on mental causation:

“Nonreductive physicalists endorse the principle of mental causation, according to which some events have mental causes: Sid climbs the hill because he wants to. Nonreductive physicalists also endorse the principle of physical causal completeness, according to which physical events have sufficient physical causes: Sid climbs the hill because a complex neural process in his brain triggered his climbing.” (Moore, 2019, p. 479)

What would it take for an end-involving scientific account of Sid’s behavior? An attractive option has been offered by dynamical approaches, which would broadly claim that Sid’s intention is either supervenient on—or even identical with—the organizational constraints that collective dynamics impose on neural or organismic activities (Freeman, 1999; Juarrero, 1999; Kelso, 1995; Thompson & Varela, 2001). Traditionally, the focus has been on collective dynamics in the brain and on mind-brain identity, but embodied versions where mind is identified more broadly with organism-environment interaction dynamics are also conceivable (Myin & Zahnoun, 2018). According to these dynamical approaches, if there were no empirical evidence of the right kind of organizational constraints operating on Sid’s organism-environment processes, then this would indicate the absence of end-involvement in his behavior. We could conceive of a situation in which Sid climbs the hills but without the presence of such internal constraints, for example because his bodily movements are externally forced by means of a robotic exoskeleton.

Yet dynamical approaches raise the worry that they only capture formal properties (Vial & Cornejo, 2022). Specifically, if the organizational constraints are sufficient causes for the changes in organismic activity, then what role do the ends themselves play in bringing about these changes? Put differently, if physiological constraints do all the actual work in getting Sid’s behavior appropriately organized, his ends as such become superfluous; they might as well be non-existent. The irony of this approach to teleology, e.g., aiming to accommodate a realist interpretation of intentions by recasting their role in terms of a non-intentional cause, is that it is self-undermining (Cae, 2023). In other words, any account of teleology’s role in generating a behavior that completely identifies this role with a physiological cause, which is itself already sufficiently accounted for by past non-teleological causes, becomes vulnerable to argument’s for an exclusion of teleology’s role in behavior generation altogether. Related concerns about efficacy can be raised regarding the role of meaning in AI (Froese & Taguchi, 2019), and the role of conscious experience in embodied action (Froese & Sykes, in press).

Accordingly, we need a stricter criterion that places a double demand on an explanation of end-involvement: it is not only the case that end-involvement in bodily processes must potentially be empirically distinguishable from its absence, but this distinguishability cannot be based on bodily events that already have sufficient causes in their own, non-teleoglical domain, thereby avoiding exclusion of end-involvement under conditions of strict physical causal completeness in that domain (Kim, 2005; Moore, 2019). We capture this stronger criterion of end-involvement in bodily processes by proposing a Participation Criterion:

Participation Criterion: *End-involvement in a bodily process entails that, in principle, it is distinguishable from one without end-involvement, specifically in terms of physiologically unpredictable changes in unexplainable variability*.

The appeal to indeterminacy has the double advantage of specifically avoiding charges of causal exclusion due to overdetermination (Potter & Mitchell, 2022), while at the same time allowing for the possibility that there is an additional factor making a difference to the bodily process, albeit a factor that is not directly intelligible at that particular scale and form of observation.

It is important to highlight that the Participation Criterion is not so strong that it would be ruled out by an inconsistency with physics. Quantum mechanics implies that when we conduct a measurement, the “particle’s response (to be pedantic – the universe’s response near the particle) is not determined by the entire previous history of the universe” (Conway & Kochen, 2009). This indeterminism could conceivably be amplified across all scales of the organism: “In general, physics is non-linear and large effects of small changes are well known to happen. From this perspective, agency is simply a situation where scale separation does not hold: nothing puzzling here” (Rovelli, 2021). There are also various bodily factors contributing to an organism’s rate of entropy production (REP) across its multi-scalar organization, which become difficult to disentangle in practice: “if one measures the total entropy production of an organism, what changes in REP are owed to the focal behaviour (e.g. locomotion) and what to the other processes playing out at different scales (e.g. perception, motor control, metabolism)?” (De Bari, Dixon, Kondepudi, & Vaidya, 2023, p. 18). This unexplained variability across all bodily scales provides a window of opportunity to satisfy the Participation Criterion.

In the next sections, we will examine the Participation Criterion through the lens of two thermodynamic accounts of end-involvement (Deacon & García-Valdecasas, 2023; Swenson, 2023), which each provide important insights, namely into the role of energy flows and autonomous organization, respectively. Nevertheless, they are shown to be insufficient in their current formulations. We will then introduce “irruption theory” (Froese, 2023) to illustrate how an end-involving account could build on these insights while better satisfying the Participation Criterion. These considerations point toward a clarified theory of the organism, in which matter, life, and mind are interdependent yet distinct phenomena that play unique roles.

1. **Energy flow: Is entropy’s “striving” a foundation for unification?**

Ecological psychology has a laudable history of developing embodied approaches to the study of life and mind. It has not only excelled in its experimental research program on perception, but also pioneered an innovative theoretical perspective on the thermodynamics of the organism (Swenson & Turvey, 1991). Compared to other systems theoretic approaches that conceptualize the organism as a self-producing system but do so in much more abstract terms, such as traditional autopoietic theory (Ruiz-Mirazo & Moreno, 2004), ecological psychology's concept of an “autocatakinetic” (ACK) system is distinctive for its appeals to actual physical phenomena (Chemero, 2012). It is deeply rooted in the thermodynamics of dissipative structures (Prigogine, 1997), often referred to as flow structures. Further, its hypothesis regarding the lawful origins of ACK systems is underpinned by the so-called “maximum entropy production principle” (MEPP), also sometimes called the “law of maximum entropy production” (LMEP), or, even more ambitiously, the “fourth law of thermodynamics” (Beretta, 2020; Morel & Fleck, 2006; Swenson, 2009, 2020). In a recent publication, Swenson (2023), a principal contributor to what he refers to as the “ACK-LMEP” paradigm, further raised the stakes, by positing that it amounts to “a grand unified theory for the unification of physics, life, information and cognition (mind)”.

Yet despite ecological psychology’s appeals to entropy production’s lawlike nature, there are open questions about the epistemological status of the MEPP (Sánchez-Cañizares, 2023), including how universally applicable the MEPP is to non-living and living systems (De Bari et al., 2023), although mathematical progress is being made (Endres, 2017). More importantly for current purposes, concerns had been raised about the adequacy of straightforward generalizations of this paradigm to living beings (Barrett, 2020b; Froese, Weber, Shpurov, & Ikegami, 2023). There is a worry that the MEPP, including Swenson’s ACK-LMEP paradigm, does not have the required conceptual resources to make sense of the end-directedness that distinguishes the motivated activity of living, cognitive ACKs from the non-motivated activity of other, generic ACK systems, for example a Benard cell (BC). Swenson dismisses these concerns as unfounded (Swenson, 2020), but this may rest on a misunderstanding of the nature of these concerns (Barrett, 2020a). Indeed, Swenson’s (2023) attempt to generalize ACK-LMEP into a grand unified theory has usefully brought the same concerns to the forefront.

To anchor our discussion, let us refer to a standard definition of ACK systems:

“ACKs are flow structures, their identities constituted through flow, and defined as

a system that maintains its ‘self’ as an entity constituted by and empirically traceable to a set of nonlinear (circularly causal) relations (constitutive relations) through the dissipation or breakdown of environmental potentials (resources) in the continuous coordinated motion of its components [4,5,7,8].

From the definition, it is simple to see that all living systems from cells to ecosystems at whatever scale (including the planet itself) are ACK systems as are abiotic systems such as dust devils, tornadoes and experimental systems like the BC.” (Swenson, 2023, p. 8)

The next step is to account for the lawful origin of ACK systems in a couple of steps of logical reasoning, which involves positing the LMEP as a general physical selection principle that can account for this spontaneous ordering:

“Specifically, ‘which paths out of available paths will a system take to get to equilibrium (maximize the entropy or minimize potentials)?’ The second law, of course, is mute on the subject. It only says that in all natural processes the entropy increases. The answer to this question, and the one that solves the entire question of physical selection, the ‘why’ of universal ordering, life and cognition is the law of maximum entropy production (LMEP) […]

(the world) a system will select the path or assembly of paths out of available paths that minimizes the potential or maximizes the entropy at the fastest possible rate given the constraints”

At this point, Swenson builds on work by Schrödinger, von Bertalanffy, and Prigogine, arguing that ordered systems tend to have a faster rate of entropy production. Essentially, the idea is that order production by a non-equilibrium system must necessarily be compensated by at least an equivalent amount of disorder production, in line with the balance equation of the Second Law, *ΔS* > 0. This brings the generation of order into the remit of the LMEP. Essentially, Swenson’s argument is that out-of-equilibrium systems will spontaneously become more organized, for example self-organizing into ACK systems, to the extent that this increase in order has the immediate consequence of an increase in entropy production due to increased efficiency of energy dissipation.

This general insight is starting to be more widely recognized as relevant for understanding the organization of life and mind (Azarian, 2022; Schneider & Sagan, 2005). However, we must proceed carefully in moving from non-living to living systems. As Swenson highlights: “All living/cognitive systems are ACK systems, but not all ACK systems are living or cognizing.” (ibid., p. 11). An essential difference is that, contrary to the MEPP, it is not in the best interest of living ACK systems to dissipate potential energy at the fastest possible rate, especially giving that this would entail getting to thermodynamic equilibrium with the environment at the fastest possible rate, which is equivalent to death (Deacon & García-Valdecasas, 2023). In recognition of this, Swenson admits that a distinctive characteristic of living ACK systems is their capacity to resist fastest *local* dissipation by being directed at spatiotemporally *distant* ends:

“This, the intentionality of living things, is life’s central distinguishing feature. Living systems are epistemic (cognitive) systems that constitute their ACK over times and distances that are arbitrary with respect to local potentials using instead their ‘on board’ potential … and information (in the semantic or meaningful sense)3 to seek out and access non-local potentials and access otherwise inaccessible dimensions of space–time [5]. The dramatic increase to otherwise inaccessible dissipative dimensions afforded by the origin and progressive ordering of life and its cognitive functioning answers the ‘why’ question in the specific case.” (Swenson, 2023, p. 12)

There is a lot to unpack in this paragraph, and several argumentative leaps are made that require careful consideration.

First, the laws of thermodynamics do not have foresight, and so nature can only ‘select’ from among the paths that are locally available to it. In other words, the MEPP is spatiotemporally constrained to competing gradients in the here and now. A key unsolved issue in this regard is how to even determine the spatiotemporal scale or system boundary with respect to which maximum entropy production is defined (Sánchez-Cañizares, 2023; Virgo, 2010). Assuming that this fundamental issue can be solved for the case of a living system, an attractive idea is that local paths can be adaptively changed by investing stored up energy to create alternative potential energy gradients with better future prospects, which can then get ‘selected’ because they dissipate potential in the fastest manner (Tschacher & Haken, 2007).

Second, this pushes back the original problem to another unsolved problem, namely the origins of stored energy potential. If the MEPP is assumed to be the driving principle behind the origins and progressive ordering of ACKs, then the sequence of thermodynamically allowed paths from a generic ACK system to the first living ACK systems must have been via paths of consistently increased rates of entropy production. Such a direct rate-dependent pathway from non-living to living does not seem plausible. Life requires something else. As Swenson acknowledges: “living systems behave arbitrarily with respect to their local potentials” (ibid., p. 12). But does life’s rate-*independent* behavior not imply by definition that there are alternative rate-*dependent* paths with faster rates of energy dissipation available? If so, then this would seem to rule out the possibility that the paths that led from non-living to living ACKs, and hence from rate-dependent to rate-independent paths, were brought about lawfully by the MEPP. At least, we are owed an account of the origins and role of this arbitrariness in the process.

An example of a rate-independent constraint on behavior mentioned by Swenson is the genetic system. However, the complexity of the genetic code is sufficiently high that it is unlikely to have arisen by chance, and hence selection by evolutionary or proto-cellular processes is typically invoked (Froese, Campos, Fujishima, Kiga, & Virgo, 2018). So far, the ACK-LMEP paradigm does not offer an alternative account of the origins of this kind of rate-independent constraint on behavior. More generally, it does not seem permissible to appeal to the potential of future increases in energy dissipation to account for the actual path selection in the present, even if these paths eventually do lead to a more efficient organization. As such, the ACK-LMEP framework does not (yet) have the conceptual resources that are required to become a grand unified theory of matter, life, and mind.

This lack of conceptual resources is also evident when considering whether the Participation Criterion is satisfied. Where is the locus of agency in the ACK-LMEP framework? It seems that it comes from a rather unexpected place, as indicated by Swenson’s insistence on the original notion of “striving” attributed to the Second Law: “‘The universe,’ Clausius [11] wrote (in an often misquoted phrase), ‘strives (*strebt*) to increase its entropy to a maximum’.” In accordance with this teleological interpretation, there is an experimental research program in ecological psychology that attempts to ground the striving of organisms in the assumed end-directedness of the general class of dissipative structures (De Bari et al., 2023). Swenson accepts this pan-teleological account of entropy production:

“It gives us the agential or motivated activity missing in theories of cognition (mind) and its evolution including the most recent 4E post-computationalist theories [27,35] whose Cartesian dualist ancestry continues to bleed through. But while LMEP gives us ACK systems, and ACKLMEP most significantly thus gives us the urgency towards existence, the filling out of space–time and the generics of spontaneous ordering, it does not give us the special cases of life and cognition.” (Swenson, 2023, p. 11)

In other words, on this view, a striving existence is already a property of non-living and non-cognitive ACK systems, because it is inherent in the universal entropic tendencies supposedly captured by the MEPP. For a framework that is at pains to distance itself from the dualistic origins of modern science, it is strange that it seems to oscillate between two arguably even less desirable ontologies: nihilism (we reduce our end-directedness to nothing but entropic tendencies) or panpsychism (we elevate entropic tendencies to striving fundamentally akin to our own end-directedness). The intuition that thermodynamically maximizing tendencies of dissipative structures have a role to play as an organizing principle in accounts of end-directed behavior seems reasonable, but this ambiguous relationship to two equally consistent overgeneralizations – either teleology is nowhere or it is everywhere – reveals that the ACK-LMEP paradigm in its current formulation fails to satisfy the Participation Criterion.

1. **Individuation: Putting constraints on entropy production**

What ecological psychology needs to get clear on is how it is possible for a dissipative structure to attain the behavioral flexibility of organisms. As a starting point, it needs to be able to provide an answer to the question: how is it possible for a dissipative structure to down-regulate its rate of energy dissipation, which would at the same time permit it to become less dependent on the dissipation of local energy potentials? This question highlights a deep and unresolved tension between the physical tendency of rate maximization and the biological capacity for rate regulation. In the absence of this regulatory capacity, the MEPP becomes self-undermining:

“This leaves us with a conundrum. In order to generate and maintain organization, living processes must take advantage of self-organizing processes, and yet they must also prevent these processes from depleting the very gradients that drive them. So, how can life both use self-organization at the same time that it prevents or holds off its terminal tendencies?” (Deacon & García-Valdecasas, 2023, p. 8)

The experimental dissipative structures currently investigated by ecological psychology do not yet lend themselves to addressing this question. But in the meantime, this gap might be filled by other theoretical models, for example of a minimal living system that consists of interdependent processes of metabolic self-production and self-other boundary generation. A line of theoretical research that has developed this possibility is Deacon’s “autogen” model (Deacon, 2012, 2021), which consists of two inderdependent processes, namely reciprocal catalysis and self-assembly:

“In other words, each of these self-organizing terminal processes—reciprocal catalysis and capsid shell self-assembly—generates the boundary conditions that the other requires, but in addition prevents the other from reaching an irreversible terminal state. As a result, the synergistic coupling of both processes will develop toward a target state that, although relatively inert, preserves the potential for both self-organizing capacities to recur when conditions are right. This targeted disposition is teleological (i.e. future-oriented).” (Deacon & García-Valdecasas, 2023, pp. 10-11)

Deacon (2021) maintains that such a system additionally exhibits individuation, autonomy, and normativity, among other system-level properties. Accordingly, the autogen model is akin to a concrete instantiation of an adaptive autopoietic system: it is a self-producing network of processes, constituting a systemic identity, which is predisposed to avert tendencies that would otherwise take it beyond its boundary of viability (Di Paolo, 2005). This intrinsic individuation is another essential ingredient for considerations of end-involvement, which goes beyond generic non-equilibrium energy flows. In the case of a physically self-producing system, the conditions for its existence and behavior are internally interdependent, and hence there is the possibility of viability-based behavior (Egbert et al., 2023).

Another advantage of the autogen model is that, by reciprocally counteracting the inherent tendency of physical processes to run down at the fastest possible rate, the problem of the whole structure’s tendency for uncontrollable dissipation of local energy potentials has been avoided. However, the solution to the problem of rate control has its problem: a system with an inherent disposition to become essentially inert, unless it is externally forced to behave, seems far removed from the class of intrinsically active living systems and their outward-reaching behavioral ends (Froese, 2021). In other words, the tendency for rate maximization was averted only at the cost of replacing it with another extremal tendency, namely the minimization of dissipation of energy potentials until it is absent altogether – complete stasis. We went from one thermodynamic extreme to another – from maximum to minimum flow – both of which are tendencies that by themselves fail to capture the flexibility of the living. It is no surprise that Deacon thinks of the autogen model in terms of a virus, rather than a single cell organism. If so, if the category of life is sufficiently broad to include viruses, then what distinguishes organisms?

Intriguingly, the autogen model fails to satisfy the Participation Principle by staying at the level of biochemistry while leaving no role for the involvement of end-directedness as such. We can describe the self-maintaining trend toward stasis as being teleologically “future-oriented”, but this is just a convenient heuristic. The processes that govern the autogen are fully specified in the here and now, and there is no additional factor at work. Like dynamical approaches to end-directedness more generally, Deacon’s approach therefore does not go beyond a kind of proto-cellular version of identity theory, which has well-known problems regarding physical efficacy. To be fair, least Deacon is explicit about this limitation: on his view, teleology is what he calls an “absential” phenomenon (Deacon, 2012), and absences do not have any physical efficacy in their own right (Deacon & Cashman, 2016). And indeed, no appeal to end-involvement in the autogen is needed to fully account for its behavior generation – it is life without mind.

Still, the autogen model has provided us with additional clues about what to look for in an alternative, genuinely end-involving account at the thermodynamic scale: we need an account of how a dissipative structure could gain the capacity to flexibly inhibit the rate of its extremal maximizing and minimizing processes. Ideally, this capacity for inhibition should also enable the structure to partially free its constitutive processes from being forced by the contours of local energy gradients, thereby ultimately permitting it to become responsive to nonlocal energy potentials, and hence making available new forms of behavioral complexity. In addition, when addressing shortcomings of current theory, we should maintain Deacon’s absential approach as a kind of methodological exclusion principle; we should continue to treat the involvement of teleology as the “hypothesis of last resort”, to paraphrase Sagan (Sagan, Thompson, Carlson, Gurnett, & Hord, 1993). Explaining a bodily process by appealing to end-involvement is only scientifically defensible if there is no purely physiological explanation.

1. **Irruption theory: Making mind matter**

Let us recap. We act in accordance with our intentions, but without having the experience of how our intentions are transformed into the appropriate physiological basis of our actions. At the same time, when we scientifically investigate the physiological basis of our actions, we cannot observe anything like intentions playing a role – there is purely physiological activity. It is no wonder, then, that nearly 50 years ago, Popper and Eccles opened their humanist book on the mind-body relation on a fundamentally skeptical note:

“Without pretending to be able to foresee future developments, both authors of this book think it improbable that the problem will ever be solved, in the sense that we shall really understand this relation.” (Popper & Eccles, 1977, p. VII)

The contemporary situation remains unchanged:

“Knowing that experiences are events that have potential physical description provides no illumination about how or why this should be so. For this reason *the* hard problem of consciousness should be cast as a problem about *intelligibility.* At bottom it asks, can consciousness be made intelligible in terms of something else (traditionally, purely functional or physical categories)? I argue that it cannot – not even in principle.” (Hutto, 2006, p. 48)

The same argument applies to accounts of the end-directedness of behavior. In accordance with the Participation Criterion, we can paraphrase Hutto as follows: making ends intelligible in terms of something non-normative, such as the purely dynamical or thermodynamical categories we have been considering, is simply not possible – not even in principle. This fundamental limit on the very intelligibility of end-involvement in the physiological basis of behavior presents a severe yet underrecognized challenge to most scientific approaches. But there is one theory that has taken this unintelligibility as the starting point for an alternative kind of approach: the recently proposed irruption theory takes in-principle unintelligibility of the mind-body relation at face value (Froese, 2023). As Deacon rightly highlighted, end-directedness does not show up as such in our scientific observations of the physiological basis of behavior. But arguably the physical efficacy of agent-level motivations is also not completely absent, either. Instead, and this step is essential, we need to accept and work with the fact that, while end-directed changes and physiological changes are both involved in behavior generation, they are observationally incommensurable: *hence, end-involvement in the physiological basis of a behavior must be associated with changes that are not intelligible in relation to that basis*.

If so, then we should be able to operationalize the unintelligible changes in the physiological basis of behavior resulting from such motivational involvement, which are referred to in the theory as “irruptions”. This could for example be achieved in information-theoretic terms (Froese, 2023), but also by considering their thermodynamic implications (Froese & Karelin, 2023). Importantly, this novel conceptualization of end-involvement is not vulnerable to the exclusion principle because irruptions do not have the status of “effects” in the usual sense of physical causality. Rather, irruptions are conceptually more akin to stochastic perturbations or noise introduced into the living system by an unintelligible ‘black box’, and are therefore freed from the demands of sufficient causation. Note that irruption theory is not reducing the efficacy of end-involvement to nothing but random physiological fluctuations; to the contrary, it insists that end-directedness of behavior makes a difference in its own right to its physiological basis but without thereby being reducible to physiology, and hence this “difference” must manifest to scientific observation as unexplained variability of physiological changes.

To be fair, this is a highly unusual way of conceiving of the efficacy of end-involvement, and so it is worth unpacking irruption theory in more detail as a set of smaller axioms and related theses, which in themselves are hopefully less controversial. Irruption theory starts by accepting that an agent’s motivations as such, including being directed at future ends, is efficacious:

“**Axiom 1: Motivational efficacy**. An agent’s motivations, as such, make a difference to the material basis of the agent’s behavior.” (Froese, 2023, p. 9)

Irruption theory also accepts that the difference that is made in this way to the physiological basis is not traceable to their agent-level source. As Deacon (2012) highlighted, when observing and measuring the material record, motivations are “absential” phenomena:

“**Axiom 2: Incomplete materiality**. It is impossible to measure how motivations, *as such,* make a difference to the material basis of behavior.” (Froese, 2023, p. 9)

This sets up an apparent tension between the behavioral efficacy of agent-level motivations and their absence in the physiological basis. However, instead of rejecting one of these two axioms, irruption theory introduces a third axiom that makes all three axioms mutually consistent:

“**Axiom 3: Underdetermined materiality**. An agent’s behavior is underdetermined by its material basis.” (Froese, 2023, p. 10)

Now comes the novel theoretical move that sets the stage for satisfaction of the Participation Criterion: it is to argue that the relative level of indeterminacy that is affecting the physiological basis of behavior is motivation-dependent, due to what is referred to as irruptions:

**“Working Hypothesis of Irruption Theory**: *The more an agent’s embodied activity is motivated, the less that activity is determined by its material basis*.” (Froese, 2023, p. 10)

Drawing parallels with how physics handles the presence of unobservable phenomena that still indirectly affect measurable outcomes, the theory posits a similar dynamic within the body of living beings. Irruptions could therefore be modeled in a simplified way as a random variable in the body that is coupled to a black box containing subjective phenomena that are irreducible and unintelligible in the material context of the body. This would mean that there are end-dependent changes in the physiological basis that would appear as a kind of in-principle unexplainable variability. In this way irruption theory can satisfy the Participation Criterion while remaining consistent with physical causal closure. However, it is one thing to claim that the theory satisfies this Criterion, and another to spell out more precisely how irruptions relate to the generation of adaptive behavior. For this purpose, irruption theory develops a framework of adaptive behavior centered on three theses:

**“Irruption Thesis**: The living body is organized as an *incomplete system* such that it is open to involvement of motivations via increased material underdetermination.

**Scalability Thesis**: The living body is organized as a *poised system* such that it amplifies microscopic irruptions to macroscopic fluctuations that can impact behavior.

**Attunement Thesis**: The living body is organized as an *attuned system* such that it responds to scaled up irruptions in a context-sensitive and adaptive manner.” (Froese, 2023, p. 11)

There are different ways of unpacking these theses. The Scalability Thesis assumes that the window of opportunity for irruptions is most likely located at the smallest scales, but given the “strange loop” self-referential organization of the brain and body (Hofstadter, 2007; Varela, 1984), an alternative possibility is that irruptions occur at the system-level scale. The Attunement Thesis ensures that irruptions end up giving rise to appropriate behavior, because the space of possibilities that they open is closed in accordance with the right mixture of internal and external constraints. Much existing work in embodied cognition slots in nicely here, in particular when deployed to explain attunement in the context of meta-stable dynamics of brain and behavior (Bruineberg, Seifert, Rietveld, & Kiverstein, 2021; Tognoli & Kelso, 2014).

Regarding the Irruption Thesis, a key issue is how to best measure the apparent interference in physiological processes due to end-involvement, and how to best model the efficacy of this interference. An attractive possibility is to focus on the concept of entropy: Given that entropy is a measure of disorder in a system, then irruptions could be measured in terms of an increase in entropy production. This fits well with a growing literature showing an association between cognition and broken detailed balance in brain dynamics (e.g., Lynn, Cornblath, Papadopoulos, Bertolero, & Bassett, 2021). Relatedly, there is a tradition of artificial life modeling that demonstrates how the basis for adaptive behavior can be simulated by stochastic breaks in system dynamics (e.g., Ikegami & Suzuki, 2008). Irruption theory’s contribution to this existing research is to provide an a priori justification for why the onset of end-involvement is ideally measured and modeled by the bursts of unconstrained and unpredictable state changes.

An application of the Irruption Thesis to the thermodynamic scale is promising but remains speculative (Froese & Karelin, 2023). Here we can offer only a brief sketch of this possibility. In the context of an ACK-LMEP or autogen model, increased end-involvement in the system’s internal processes could be equivalent to increased background noise levels, which would in turn lead to a decreased availability of free energy for the system to do work. At first sight this efficacy of end-involvement as disordering factor might seem counterproductive, but it depends on the context. As we saw, the ACK-LMEP paradigm was missing a mechanism for the inhibition of physical tendencies toward maximum energy dissipation, for which the autogen model overcompensated by introducing a tendency toward minimum energy dissipation. Irruptions could provide a minimal living system with an end-dependent manner of down-regulating the efficiency at which potential energy is dissipated, thereby effectively slowing them down.

This appeal to thermodynamic inhibition as the primary consequence of end-involvement is consistent with the primordial end of life, namely survival as the “mother-value of all values” (Jonas, 1992). At the origins of life, the first end that was required was preventing the system to cross its metabolic boundary of viability, and hence regulatory inhibition of spontaneous thermodynamic tendencies would have been a fitting adaptive response. Moreover, inhibition continues to be the default mechanism of regulation even for more complex forms of life (Jost, 2021). Irruption theory could therefore be elaborated in future work to contribute to a more systematic thermodynamic grounding of current enactive accounts of adaptivity (Di Paolo, 2018). For example, sufficiently large irruptions could also serve to “reset” the living system’s state more generally by temporarily flattening the attractor landscape, thereby broadening its exploration of state space, which in conjunction with basic associative memory can facilitate self-optimization of constraints via a mechanism akin to generalization (Froese et al., 2023).

1. **Conclusion**

We have analyzed three theories of end-directed behavior that can speak to its thermodynamic basis, and we have found essential ingredients in each of them. The LMEP-ACK paradigm has demonstrated that we can get efficient energy flow from physics alone, while the autogen model highlights the role of biological organization in taming those physical tendencies in a global-to-local or system-to-component manner, and irruption theory introduces the possibility of flexible end-directed adaptive regulation by injecting arbitrariness into state dynamics. Taken together, the LMEP-ACK paradigm, the autogen model, and irruption theory highlight the complementary roles of (1) energy flow maintenance, (2) systemic identity creation, and (3) state constraint destruction, respectively. This suggests a threefold division of the ontology of the organism into characteristically distinctive domains of matter, life, and mind.

**Acknowledgement**

This research was initiated while TI was visiting the Okinawa Institute of Science and Technology (OIST) through the Theoretical Sciences Visiting Program (TSVP).

**Statements and Declarations**

No funding was received to assist with the preparation of this manuscript.

The authors have no relevant financial or non-financial interests to disclose.

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

**References**

Azarian, B. (2022). *The Romance of Reality: How the Universe Organizes Itself to Create Life, Consciousness, and Cosmic Complexity*. Dallas, TX: BenBella Books.

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.

Beretta, G. P. (2020). The fourth law of thermodynamics: Steepest entropy ascent. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences, 378*, 20190168. doi:10.1098/rsta.2019.0168

Bruineberg, J., Seifert, L., Rietveld, E., & Kiverstein, J. (2021). Metastable attunement and real-life skilled behavior. *Synthese, 199*, 12819-12842.

Cae, I. (2023). On motivating irruptions: The need for a multilevel approach at the interface between life and min. *Adaptive Behavior*. doi:10.1177/10597123231184651

Carroll, S. M. (2022). The Quantum Field Theory on Which the Everyday World Supervenes. In S. Ioannidis, G. Vishne, M. Hemmo, & O. Shenker (Eds.), *Levels of Reality in Science and Philosophy: Re-Examining the Multi-level Structure of Reality* (pp. 27-46). Charn, Switzerland: Springer.

Chemero, A. (2012). Modeling self-organization with nonwellfounded set theory. *Ecological Psychology, 24*, 46-59.

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

De Bari, B., Dixon, J., Kondepudi, D., & Vaidya, A. (2023). Thermodynamics, organisms and behaviour. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences, 381*(20220278). doi:10.1098/rsta.2022.0278

Deacon, T. W. (2012). *Incomplete Nature: How Mind Emerged from Matter*. New York, NY: W. W. Norton & Company.

Deacon, T. W. (2021). How molecules became signs. *Biosemiotics*.

Deacon, T. W., & Cashman, T. (2016). Steps to a metaphysics of incompleteness. *Theology and Science, 14*(4), 401-429.

Deacon, T. W., & García-Valdecasas, M. (2023). A thermodynamic basis for teleological causality. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences, 381*(20220282). doi:10.1098/rsta.2022.0282

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences, 4*(4), 429-452.

Di Paolo, E. A. (2018). The enactive conception of life. In A. Newell, S. Gallagher, & L. De Bruin (Eds.), *The Oxford Handbook of 4E Cognition* (pp. 71-94). Oxford, UK: Oxford University Press.

Egbert, M. D., Hanczyc, M. M., Harvey, I., Virgo, N., Parke, E. C., Froese, T., . . . Bartlett, S. (2023). Behaviour and the origin of organisms. *Origins of Life and Evolution of Biospheres*. doi:10.1007/s11084-023-09635-0

Endres, R. G. (2017). Entropy production selects nonequilibrium states in multistable systems. *Scientific Reports, 7*, 14437. doi:10.1038/s41598-017-14485-8

Freeman, W. J. (1999). *How Brains Make Up Their Minds*. London, UK: Weidenfeld & Nicolson.

Froese, T. (2021). To understand the origin of life we must first understand the role of normativity. *Biosemiotics, 14*, 657-663.

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., Campos, J. I., Fujishima, K., Kiga, D., & Virgo, N. (2018). Horizontal transfer of code fragments between protocells can explain the origins of the genetic code without vertical descent. *Scientific Reports, 8*, 3532. doi:10.1038/s41598-018-21973-y

Froese, T., & Karelin, G. (2023). The enactive account of motivated activity and the hard problem of efficacy (HPE): Artificial life meets the physics of life. In H. Iizuka, K. Suzuki, R. Uno, L. Damiano, N. Spychala, M. Aguilera, E. Izquierdo, R. Suzuki, & M. Baltieri (Eds.), *Proceedings of the Artificial Life Conference 2023 (ALIFE 2023)*. Cambridge, MA: The MIT Press.

Froese, T., & Sykes, J. J. (in press). The pragmatics, embodiment, and efficacy of lived experience: Assessing the core tenets of Varela’s neurophenomenology. *Journal of Consciousness Studies, 30*(11-12). Retrieved from <http://philsci-archive.pitt.edu/id/eprint/21970>

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. *BioSystems, 230*(104959). doi:10.1101/2023.02.05.527213

Hofstadter, D. (2007). *I Am A Strange Loop*: Basic Books.

Hutto, D. D. (2006). Impossible problems and careful expositions: Reply to Myin and De Nul. In R. Menary (Ed.), *Radical Enactivism: Intentionality, Phenomenology, and Narrative: Focus on the Philosophy of Daniel D. Hutto* (pp. 45-64). Amsterdam, the Netherlands: John Benjamins Publishing.

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

Jonas, H. (1992). The burden and blessing of mortality. *The Hastings Center Report, 22*(1), 34-40.

Jost, J. (2021). Biology, geometry and information. *Theory in Biosciences*.

Juarrero, A. (1999). *Dynamics in Action: Intentional Behavior as a Complex System*. Cambridge, MA: The MIT Press.

Kelso, J. A. S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge, MA: The MIT Press.

Kim, J. (2005). *Physicalism, Or Something Near Enough*. Princeton, NJ: Princeton University Press.

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

Mitchell, K. J. (2023). *Free Agents: How Evolution Gave Us Free Will*: Princeton University Press.

Moore, D. (2019). Causal exclusion and physical causal completeness. *Dialectica, 73*(4), 479-505.

Morel, R. E., & Fleck, G. (2006). A fourth law of thermodynamics. *Chemistry, 15*(4), 305-310.

Musser, G. (2023). *Putting Ourselves Back into the Equation: Why Physicists Are Studying Human Consciousness and AI to Unravel the Mysteries of the Universe*: Farrar, Straus and Giroux.

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

Nicolescu, B. (2012). Transdisciplinarity: the hidden third, between the subject and the object. *Human and Social Studies, 1*(1), 13-28.

Noble, D., & Ellis, G. (2022). Biological relativity revisited: The pre-eminent role of values. *Theoretical Biology Forum, 115*(1/2), 45-69.

Popper, K. R., & Eccles, J. C. (1977). *The Self and Its Brain: An Argument for Interactionism*: Springer.

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

Prigogine, I. (1997). *The End of Certainty: Time, Chaos, and the New Laws of Nature*. New York, NY: The Free 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.

Ruiz-Mirazo, K., & Moreno, A. (2004). Basic autonomy as a fundamental step in the synthesis of life. *Artificial Life, 10*(3), 235-259.

Sagan, C., Thompson, W. R., Carlson, R., Gurnett, D., & Hord, C. (1993). A search for life on Earth from the Galileo spacecraft. *Nature, 365*(6448), 715-721. doi:10.1038/365715a0

Sánchez-Cañizares, J. (2023). Is the maximum entropy production just a heuristic principle? Metaphysics on natural determination. *Synthese, 201*, 121.

Sapolsky, R. M. (2023). *Determined: A Science of Life Without Free Will*: Penguin Press.

Schneider, E. D., & Sagan, D. (2005). *Into the Cool: Energy Flow, Thermodynamics, and Life*. Chicago, IL: University of Chicago Press.

Swenson, R. (2009). The Fourth Law of Thermodynamics or the Law of Maximum Entropy Production (LMEP). *Chemistry, 18*(5), 333-339.

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. (2023). A grand unified theory for the unification of physics, life, information and cognition (mind). *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences, 381*(20220277). doi:10.1098/rsta.2022.0277

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

Thompson, E., & Varela, F. J. (2001). Radical embodiment: Neural dynamics and consciousness. *Trends in Cognitive Sciences, 5*(10), 418-425.

Tognoli, E., & Kelso, J. A. S. (2014). The metastable brain. *Neuron, 81*, 35-48.

Tschacher, W., & Haken, H. (2007). Intentionality in non-equilibrium systems? The functional aspects of self-organized pattern formation. *New Ideas in Psychology*, 1-15.

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. (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.

Vial, I., & Cornejo, C. (2022). Not complex enough for complexity: Some intricacies of interpersonal synergies theory. *New Ideas in Psychology, 64*(100914). doi:10.1016/j.newideapsych.2021.100914

Virgo, N. (2010). From maximum entropy to maximum entropy production: A new approach. *Entropy, 12*, 107-126. doi:10.3390/e12010107

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

Watson, D. L. (1930). Entropy and organization. *Science, 72*(1861), 220-222.

Wiener, N. (1961). *Cybernetics: Or Control and Communication in the Animal and the Machine* (Second ed.). Cambridge, MA: The MIT Press.

1. I am grateful to Steve Esser for this precise formulation of the fundamental problem. [↑](#footnote-ref-1)