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

A central ambition of an expanded theoretical biology is to provide an account of how both physiology and agency, and each in their own irreducible way, contribute to the generation of adaptive behavior. To ensure that the semiotic, communicative, representational, or meaning-bearing aspects of agency make a measurable difference to behavior generation, we introduce a test for candidate theories, the "participation criterion": *End-directedness of a behavior entails that, in principle, it is distinguishable by measurement from one without end-directedness.* Two promising theories of the thermodynamic basis of end-directedness, namely Swenson's "law of maximum entropy production (LMEP)" and Deacon's "autogen" model, fall short arguably by construction. We then appeal to the realist and non-reductive "irruption theory" of agency as a compelling way forward. We speculate that end-directedness will show up in measurement as a local increase in unpredictability of physiological dynamics, which has the global effect of stochastically nudging the organism to the end state. Accordingly, irruption theory satisfies the participation criterion by predicting an end-directedness-dependent acceleration of the rate of entropy production. This prediction is consistent with existing research into the association between neural fluctuations and task behavior, is open to further experimental verification, and provides a novel perspective on the sources of entropy production in the organism.

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

1. Introduction

One of the most pressing frontiers of science is to provide a theory of how intentional agents like ourselves fit into the natural order described by the physical and life sciences (e.g. Azarian, 2022; Ball, 2023; Mitchell, 2023; Musser, 2023; Sapolsky, 2023). For example, long-standing

questions regarding the possibility of teleology, which denotes the goal- or end-directedness of biological processes, become even more intricate when considering the intentional actions of humans (Noble & Ellis, 2022). We all have first-person experiences of purposeful action, in which there is a felt disposition to achieve an anticipated end alongside a corresponding sense of agency. Yet this familiar human form of teleological causality can be considered a special case of the general form of end-directed behavior with which theoretical biology has long struggled (García-Valdecasas & Deacon, 2024). It is therefore helpful to adopt a minimalist approach with the aim of systematically uncovering the necessary and sufficient conditions of teleological causality in organisms (García-Valdecasas, 2022). Without a workable account of basic teleology, more specific claims regarding normative aspects of biological processes, including semiotic, communicative, representational, or meaning-bearing aspects, will rest on shaky foundations.

There are valuable attempts at addressing this challenge from the bottom up, for example by elucidating the end-directedness of living systems in thermodynamic terms (e.g., Deacon, 2012; Swenson & Turvey, 1991; Tschacher & Haken, 2007). Yet, akin to the causal exclusion principle that is haunting nonreductive physicalism more generally (Kim, 2005), these proposals may still suffer from a notable theoretical shortcoming: the assumed presence of end-directedness in these minimal systems is, empirically, indistinct from its absence[1]. For example, if we accept that ends supervene on biological processes in a physical world characterized by causal completeness at that scale, then appeals to end-directedness in behavior generation can at best be a heuristic tool. There would be no conceptual room for the efficacy of end-directedness as such. But if the presence of teleology is compatible with its absence, then is equivalent to treating teleology as physically ineffective, which is hard to square with the appearance of end-directedness across all biological scales (Ball, 2023). It would also ultimately be in tension with our own sense of agency, including as scientists pursuing ends by developing theories.

Accordingly, the challenge is to develop a scientifically workable theory of the efficacy of end-directedness, which also does justice to the first-person experience that the presence of end-directedness genuinely makes a difference to behavior. We can reformulate this challenge into an explicit criterion to assess how successful a theory of behavior generation is in securing the efficacy of end-directedness. 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 is the role of end-directedness in Sid's behavior? A loosely related group of dynamical approaches 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

[1] TF is grateful to Stephen Esser for this precise formulation of the fundamental problem.

activities (Deacon, 2012; Freeman, 1999; Juarrero, 1999; Kelso, 1995; Thompson & Varela, 2001). Traditionally, the focus has been on the brain and on mind-brain identity, but embodied versions where mental features are identified more broadly with organism-environment interaction dynamics are also conceivable (Myin & Zahnoun, 2018). Accordingly, if there were no evidence of appropriate organizational constraints on Sid's bodily processes, then this would indicate the absence of end-directedness in his behavior.

Yet this focus on organizational constraints raises the worry that dynamical approaches capture only formal properties (Vial & Cornejo, 2022). Specifically, if the organizational constraints were sufficient causes for changes in organismic activity, then what role would the ends themselves play in bringing about these changes? Put differently, if these constraints do all the actual work in getting Sid's behavior appropriately organized, his ends as such become superfluous. Or at least, the normative conditions associated with ends, but not with mere constraints, would not make any difference – the efficacy would rest on the constraints alone. In other words, the properties that make ends distinctive from generic constraints might as well be non-existent. The irony of such an approach to naturalizing teleology, e.g., of aiming to accommodate a realist interpretation of intentions by recasting their role purely in terms of a non-intentional cause, is that it is self-undermining (Cae, 2023).

More generally, any theory of teleology's role in behavior generation that identifies that role with a concrete physiological process or cause is one step away from eliminativism. To put it differently, if a theory appeals to a particular physiological factor as the role that end-directedness is supposed to play in the physiology of behavior generation, it can justifiably be asked whether that end-directedness as such does any work itself, or whether it is ultimately nothing but an ineffective epiphenomenon. Related concerns about how to make proper room for the efficacy of mental features can be raised regarding the role of meaning in AI (Froese & Taguchi, 2019), and the role of lived experience in embodied action (Froese & Sykes, 2023).

Accordingly, we need a stricter criterion that places a two-fold demand on any explanation of end-directedness: it is not only the case that end-directedness of behavior must potentially be empirically distinguishable from its absence, but this distinguishability cannot be based on physiological aspects that already have a sufficient account in their own terms. It is in meeting the second demand that a theory can manage to avoid the exclusion of end-directedness even under the strict assumption of a physical causal completeness, as posited by conservative naturalists (Kim, 2005; Moore, 2019). We capture this stronger demand for an irreducible role of end-directedness in behavior by proposing a Participation Criterion:

> Participation Criterion: *End-directedness of a behavior entails that, in principle, it is distinguishable by measurement from one without end-directedness.*

Foreshadowing the proposal we will develop in more detail in section 4, we briefly note that one scientifically workable strategy for satisfying the Participation Criterion is to conceptualize the efficacy of end-directedness in behavior generation in terms of a spontaneous change in stochastic fluctuations. This kind of 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 an additional factor dependent on end-directedness is making a measurable difference to the bodily process, albeit a factor that is itself not directly observable from that particular measurement perspective.

It is important to preemptively address the reasonable concern that the Participation Criterion is too strong, such that it would be ruled out because of inconsistency with physics. Fortunately, there is sufficient causal slack in organisms' behavior. At least on some interpretations of quantum mechanics, when observers conduct a measurement, the result is not determined by the previous state of the universe (Conway & Kochen, 2009). And 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).

It is also noteworthy that just because a theory satisfies the Participation Criterion in principle, this does not mean that the difference can be easily measured in practice. The various physiological processes contributing to an organism's rate of entropy production (REP) across its multi-scalar organization are difficult to disentangle. Accordingly, De Bari et al. (2023, p. 18) ask: "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)?" Thus, the vast uncertainty inherent in biological processes provides a window of opportunity for theories of end-directedness to satisfy the Participation Criterion.

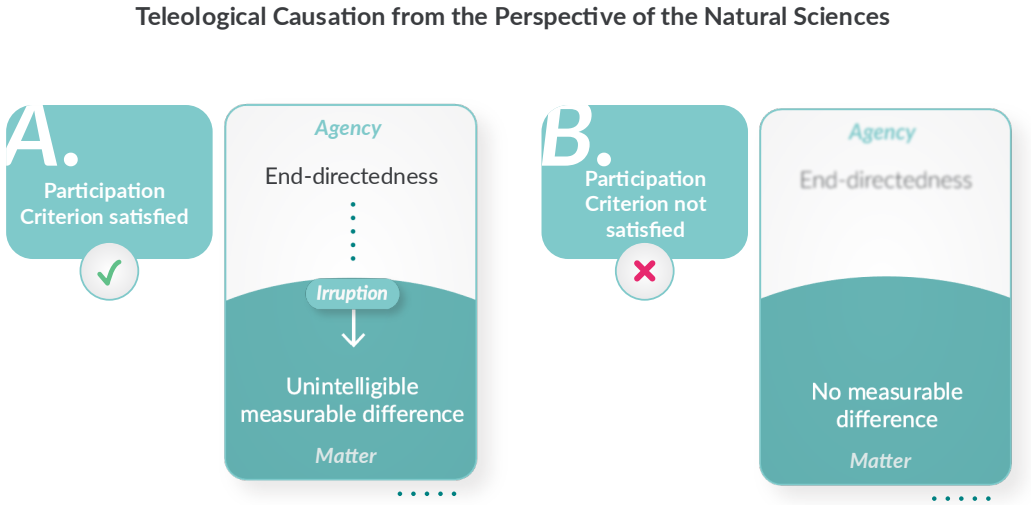

Figure 1. Illustration of the proposed Participation Criterion for theories of end-directedness. Panel A. represents the hypothesis that there can be co-dependent existence of two distinct but coupled domains, in this case end-directed agency (teleology) and living matter (physiology), that is mediated by cross-domain interaction, and which in turn entails a measurable difference. However, the interaction itself is unobservable, and hence the difference (irruption) is not intelligible from inside a single domain. In contrast, in Panel B., according to theories in which cross-domain interaction is not allowed, there can be no effects due to this interaction, and therefore no measurable difference. In that case, the domain of agency has no distinctive efficacy on the domain of physiology and could be eliminated.

In the next sections, we examine two of the most developed thermodynamic accounts of end-directedness, namely the "law of maximum entropy production" by Swenson and colleagues as well as the "autogen" model by Deacon and colleagues, through the lens of the Participation Criterion. Each of the theories provides important insights, specifically regarding the roles of energy flow and of autonomous organization, in end-directed processes. Nevertheless, they fall short of satisfying the Participation Criterion, at least in their current formulations. We will then introduce the irruption theory of intentional agency as an example of how a theory of end-directedness could build on their insights while also satisfying the Participation Criterion. Our suggestions of how the efficacy of end-directedness manifests in terms of thermodynamics, and what is its broader role in physiological activity, remain speculative. Yet they tantalizingly point toward an expanded theoretical biology, in which matter, life, and mind are three distinct yet related domains of phenomena that play unique and complementary roles in the organism.

2. Matter: Energy flow

One innovative theoretical perspective on the thermodynamics of the organism comes from ecological psychology (Swenson & Turvey, 1991). In contrast to systems-theoretic approaches that conceptualize the organism as a self-producing system in more abstract terms, such as Maturana and Varela's (1980) traditional autopoietic theory (cf. Ruiz-Mirazo & Moreno, 2004), ecological psychology's concept of an "autocatakinetic" (ACK) system is distinctive for its appeals to physics (Chemero, 2012). It is 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" ((Deacon, 2021)), 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 it as "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 and ontological 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). There has been numerical work showing that in some multi-stable systems the steady state with the highest entropy production is favored (Endres, 2017), but there are counterexamples (Bartlett & Virgo, 2016).

More importantly for our current purposes, there remains a specific concern about the adequacy of a straightforward application of the ACK-LMEP paradigm to the specific characteristics of the behavior of living beings (Barrett, 2020b; Froese, Weber, Shpurov, & Ikegami, 2023). It seems that the ACK-LMEP paradigm does not (yet) have sufficient conceptual resources to distinguish the end-directedness of living, cognitive ACKs from the non-normative processes of other, generic ACK systems, for example a Bénard cell (BC). Swenson (2020) has

dismissed this concern as unfounded, but this dismissal may rest on a misunderstanding of the criticism (Barrett, 2020a). Indeed, Swenson's (2023) subsequent attempt to turn the ACK-LMEP paradigm into a "grand unified theory" has usefully brought this same concern to the forefront.

To unpack the concern in more detail, 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 […]." (Swenson, 2023, p. 8)

In other words, ACKs are a specific class of dissipative structure, which includes both living and non-living systems from cellular to planetary scales. The next step is to account for the lawful origin of ACK systems, which involves positing the LMEP as a general selection principle the provides an answer to the question of path selection:

> "'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) or the fourth law of thermodynamics [1,4,5,26,31-33]:
>
> > (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" (Swenson, 2023, p. 10)

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. However, we must proceed carefully in moving from non-living to living systems. Contrary to the LMEP, it is not in the best interest of living ACK systems to always dissipate free energy at the fastest possible rate, especially giving that this would entail approaching thermodynamic equilibrium with the environment at the fastest possible rate, which is equivalent to dying (Deacon & García-Valdecasas, 2023). In recognition of this problem, Swenson admits that a distinctive characteristic of living ACK systems is their capacity to resist the fastest *local* dissipation by redirecting dissipation toward 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 require more careful deliberation.

To begin with, 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 LMEP 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 could 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).

However, this pushes back the original problem to another unsolved problem, namely the origins of stored energy potential. If the LMEP 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 rate-dependent pathway from non-living to living does not seem plausible. Life is distinctive, as Swenson acknowledges: "living systems behave arbitrarily with respect to their local potentials" (ibid., p. 12). An account of the origins of this arbitrariness in living ACKs is still missing.

As an example of a rate-independent constraint on behavior, Swenson refers to the genetic system. However, the genetic code is sufficiently complex that it is unlikely to have arisen by chance, and hence selection by evolutionary or proto-cellular processes is required (Froese, Campos, Fujishima, Kiga, & Virgo, 2018). And it is not permissible, at least not without invoking something akin to teleological backward causation, to appeal to future increases in energy dissipation to account for the selection of the present path. Hence, the locus of agency is directly identified with locally increasing entropy production, 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 of entropy production, there is an experimental research program in ecological psychology that attempts to ground the striving of organisms in the assumed end-directedness of dissipative structures (De Bari et al., 2023).

In sum, according to the ACK-LMEP paradigm, end-directedness is a property of all ACK systems, whether living or non-living, because it is identified with universal entropic tendencies. The system's goal just is the maximization of entropy production. Therefore, end-directedness as such no longer has any specific role of its own to play in behavior generation. This has the

benefit of satisfying a strict naturalism, yet it does so by sacrificing a nonreductive realism about end-directedness. By construction, the current formulation of the ACK-LMEP paradigm thereby fails to satisfy the Participation Criterion.

3. Life: Individuation

What the ACK-LMEP paradigm needs to get clearer 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 answer the question: how is it possible for a dissipative structure to down-regulate its rate of energy dissipation, which would involve getting a degree of independence from the dissipation of local energy potentials? This question highlights a deep and unresolved tension between the assumed universal tendency of entropy rate maximization and the biological capacity for rate regulation. In the absence of this regulatory capacity, the LMEP 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 dissipative structures that are investigated by ecological psychology cannot (yet) address this question. However, another line of theoretical research into the thermodynamics of end-directedness that has positioned itself as providing an answer is Deacon's "autogen" model (Deacon, 2012, 2021). An autogen consists of two interdependent processes, namely reciprocal catalysis and self-assembly:

> "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)

An advantage of the autogen model is that, by reciprocally counteracting the tendency of physical processes to run down, the problem of the whole structure's tendency for maximum dissipation of local energy potentials has been avoided. However, the solution raises a different concern (Froese, 2021): an autogen has a disposition to become inert, unless it is externally forced to react. In other words, a one-sided tendency was averted only at the cost of replacing it with another one-sided tendency, namely the minimization of dissipation of energy potentials until dissipation ceases altogether – complete stasis. We went from one extreme tendency to another – from maximum flow to no flow – both of which are tendencies that by themselves fail to capture the flexible behavior of living systems. As Deacon (2023) points out, this lack of a capacity to initiate behavior is by design, as it helps to simplify the autogen model. For example,

there is no need to assume that an autogen has the capacity to accumulate and store potential energy, and hence questions about the origins of this capacity can be deferred.

Still, the autogen model notably sets the bar higher for end-directedness compared to the ACK-LMEP paradigm. The latter identifies end-directedness with the self-organized increase of entropy production in a pre-existing physical system, such as an electrical dissipative structure consisting of metal beads in a fluid (De Bari et al., 2023). The autogen model is situated in the more complex domain of chemistry, in which an enclosed system self-organizes out of specific interdependent processes of catalysis and crystallization. This process is taken to be a "distinctive end-directed dynamic" (García-Valdecasas & Deacon, 2024, p. 75), but it is just a heuristic that plays no role in the dynamics of the autogen model. Nothing but chemical interactions are at work in the model. Indeed, Deacon is explicit about not assigning efficacy to end-directedness as such: teleology is part of the broader class of what he calls "absential" phenomena, whose absence from direct observation is a property that he argues facilitates their naturalization in terms of constraints (Deacon, 2012). Absential phenomena do not even have physical efficacy as constraints alone. As Deacon and Cashman (2016) clarify, doing work requires both contextual constraints and energy release:

> The "efficacy" confusion is also related to this misidentification of absence with non-being. Defining the concept of *constraint* in terms of absent degrees of freedom makes it tempting to think of absences *doing* things. But absences themselves don't do work, nor do they resist work. And yet there is no work without absence. The absent degrees of freedom are only part of the story, necessary but not sufficient. Physical work requires the release of energy in a *constrained* context. (Deacon & Cashman, 2016, pp. 419-420)

In sum, the teleological causality of the autogen model is a "physically embodied disposition" with a "material existence that can be preserved or lost" (García-Valdecasas & Deacon, 2024, p. 75). We can conclude that, like the ACK-LMEP paradigm, the autogen model by construction does not satisfy the Participation Criterion: given that the chemical dynamics of this model can be completely specified in terms of the physics of constrained energy release, the presumed presence of teleology as such makes no difference compared to its absence.

Still, the autogen model has provided us with useful clues about what to look for as we move from chemical systems to living systems: we need an account of how a dissipative structure could gain the capacity to flexibly inhibit its own tendencies. Ideally, this capacity for inhibition should enable the structure to free its processes from always being driven by local energy gradients, thereby permitting it to become responsive to nonlocal energy potentials, and hence ultimately making available new forms of behavioral complexity.

In addition, we can build on Deacon's "absential" approach and go a step further: appeals to the presence of teleology are only permissible as a "hypothesis of last resort", to paraphrase Sagan (Sagan, Thompson, Carlson, Gurnett, & Hord, 1993). Positing an efficacious role for end-directedness in behavior generation is only justifiable for those measurable differences for which an immediate physiological cause is absent or at least has not been observed. We

therefore could not agree with Deacon's claim that "a good model should include no unknown or undescribed processes" and "include no opaque (black box) properties" (Deacon, 2021, p. 541). In contrast, we believe that taking both end-directedness and its "absential" nature seriously highlights the need of broadening the scope of admissible phenomena to those that are not immediately observable.

4. **Mind: Irruption**

Let us assume that we act freely in accordance with our goals. Yet we do not have first-hand access to precisely how our goals are transformed into the appropriate physiological basis of our behavior. At the same time, when we scientifically investigate the physiological basis of behavior generation, we cannot directly measure anything like end-directedness playing a role – there is purely physiological activity. We have therefore argued that explaining end-directedness in lower-level dynamics is simply not possible – because goals do not exist on that lower level in the first place. This limit on the intelligibility of end-directedness is a severe challenge.

The recently proposed irruption theory takes this in-principle limit at face value (Froese, 2023; Froese & Karelin, 2023). As Deacon rightly highlighted, end-directedness does not show up as such in our observations of the physiological basis of behavior. At the same time, the efficacy of end-directedness cannot be completely absent at that scale, either. Instead, and this is crucial, we need to start working with the fact that, while both end-directed and physiological aspects are involved in behavior generation, only the latter are directly accessible via measurement. We are therefore led to posit the following research hypothesis:

End-directedness of a behavior is associated with measurable changes at the scale of physiology that cannot be fully predicted purely from that physiological basis alone.

If so, then we need to operationalize the changes in the physiological basis resulting from end-directedness, which are referred to in the theory as "irruptions". Irruptions are akin to variable stochastic perturbations or noise introduced into the living system by a 'black box', which stands for the efficacy of end-directedness. To be fair, this is a highly unusual way of conceiving of the efficacy of end-directedness, and so it is worth unpacking irruption theory in more detail as a set of smaller axioms and related theses, which in themselves are 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 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 with which the Participation Criterion is satisfied: the relative level of indeterminacy of the physiological basis of behavior is dependent on the presence of end-directedness, due to irruptions making a difference. In other words, there are end-directed-dependent changes at the scale of physiology in terms of stochastic variability that would be absent otherwise. However, it remains to be spelled out how these irruptions relate to the generation of appropriately end-direct behavior. For this purpose, the theory is proposes three theses (Froese, 2023, p. 11), which we adapt to the case of end-directedness:

> **Irruption Thesis**: A living system is organized as an *incomplete system* such that it is open to end-directedness via increased physiological underdetermination.
>
> **Scalability Thesis**: A living system is organized as a *poised system* such that it amplifies microscopic irruptions to macroscopic fluctuations that impact end-directed behavior.
>
> **Attunement Thesis**: A living system is organized as an *attuned system* such that it responds to scaled up irruptions in an end-directed manner.

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 give rise to appropriate behavior, because the space of possibilities that they open is then closed down in accordance with the right mixture of internal and external constraints. Much existing work in embodied cognition slots in here, such as 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 measure the interference in physiological processes due to end-directedness, and how to 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 a temporary increase in entropy production. For example, 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 in artificial life 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 research is to provide an explanation for why the onset of end-directedness can be measured and modeled by bursts of unpredictable state changes.

An application of the Irruption Thesis to the thermodynamic scale could be promising but remains speculative (Froese & Karelin, 2023). Here we can offer only a brief sketch. In the context of an ACK-LMEP or autogen model, increased end-directedness in a system's processes could be equivalent to increased noise levels. At first sight, this efficacy of end-directedness as a 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 tendencies toward the maximum rate of energy dissipation, for which the autogen model overcompensated by introducing a tendency toward the minimum rate of energy dissipation. Irruptions could provide a minimal living system with the capacity for end-directed regulation of the rate at which energy is dissipated. For example, stochastic perturbations could degrade energy sources, or decrease efficiency of work-constraint cycles, both of which will slow processes down.

This appeal to thermodynamic inhibition as the primary consequence of end-directedness is consistent with the primordial goal of life, namely self-preservation as the "mother-value of all values" (Jonas, 1992). At the origins of life, one essential goal was preventing the system to cross its metabolic boundary of viability, and hence inhibition of thermodynamic tendencies would have been an adaptive response. A more flexible regulation could then be achieved by a simple mechanism of rein control (Harvey, 2004). Moreover, inhibition continues to be the default mechanism of regulation for more complex forms of life (Jost, 2021). Yet in the context of these evolved living systems, the end-directed-dependent presence of irruptions will also have correspondingly more complex consequences, even if their immediate impact remains the same – a contribution to stochastic fluctuations. Starting with the realization of the ubiquity of 1/f noise in natural systems (Bak, 1996), it has been increasingly recognized that noise plays an essential role in the adaptive workings of the brain (e.g. Mitchell, 2023; Northoff, 2018), and in the organism more generally (e.g. Ball, 2023; Longo & Montévil, 2014; Roy & Majumdar, 2022).

Irruption theory could therefore be elaborated to contribute to more thermodynamic grounding of 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). This comes close to Mitchell's recent argument in support of a two-stage model of action selection, which he elaborates in his systematic defense of agency and free will:

> Importantly, in this model, it's not that *individual* random events at the quantum level decide what the organism does or generate new ideas. It's that the general randomness and

> thermal fluctuations cause a kind of variability in neural networks that can jostle them out of the ruts of habit and into potentially novel states. (Mitchell, 2023, p. 189)

However, as Schurger and colleagues point out, "theorists who want to identify the source of action as the agent will have to tell a story that somehow makes a case for the noisy trigger being part of or attributable to the agent" (Schurger, Hu, Pak, & Roskies, 2021, p. 566). Irruption provides a potential source for this "active modulation of randomness" (Mitchell, 2023, p. 188), which would make a measurable difference compared to the absence of end-directedness.

5. 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 ACK-LMEP paradigm has demonstrated that we can get self-organized energy flow from physics alone, while the autogen model highlights the role of codependent processes in a self-assembling chemical system, and irruption theory introduces the possibility of end-directed regulation by injecting variability. Taken together, the ACK-LMEP paradigm, the autogen model, and irruption theory highlight the complementary roles of (1) energy flow maintenance, (2) systemic constraint construction, and (3) state constraint destruction, respectively. All three roles are necessary to explain the end-directed behavior of living systems.

Arguably, this complexity is not accidental, but an essential and irreducible aspect of our own 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, the task of an expanded theoretical biology is to describe the end-directedness of behavior as an efficacious relationship between two distinct domains of phenomena, teleology and physiology, but in such a way that neither domain can be directly identified with the other. This approach requires taking seriously the possibility of irreducible cross-domain consequences, from the scale of minimal living systems to that of the human mind (Nicolescu, 2012; Wagemann, 2011). From simple inhibition to symbolic negation, we expect that irruption can become a useful concept to account for variability in behavior.

Acknowledgement

This research was initiated while TI was visiting the Okinawa Institute of Science and Technology (OIST) through the Theoretical Sciences Visiting Program (TSVP). We thank Roger Pullin, Alexander Hölken for detailed comments on earlier drafts. We are grateful to Alina Hernandez Porrello for help with the design of the figure.

References

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

Bak, P. (1996). *How Nature Works: The Science of Self-Organized Criticality*. New York: Copernicus.

Ball, P. (2023). *How Life Works: A User's Guide to the New Biology*. Chicago, IL: The University of Chicago Press.

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.

Bartlett, S., & Virgo, N. (2016). Maximum entropy produciton is not a steady state attractor for 2D fluid convection. *Entropy, 18*(12), 431. doi:10.3390/e18120431

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

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. (2023). Minimal properties of a natural semiotic system: Response to commentaries on "How molecules became signs". *Biosemiotics, 16*, 1-13.

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. (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.

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. (2023). The pragmatics, embodiment, and efficacy of lived experience: Assessing the core tenets of Varela's neurophenomenology. *Journal of Consciousness Studies, 30*(11-12), 190-213.

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

García-Valdecasas, M. (2022). On the naturalisation of teleology: Self-organisation, autopoiesis and teleodynamics. *Adaptive Behavior, 30*(2), 103-117.

García-Valdecasas, M., & Deacon, T. W. (2024). Origins of biological teleology: How constraints represent. *Synthese, 204*(75). doi:10.1007/s11229-024-04705-w

Harvey, I. (2004). Homeostasis and Rein Control: From Daisyworld to Active Perception. In J. Pollack, M. A. Bedau, P. Husbands, T. Ikegami, & R. A. Watson (Eds.), *Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems* (pp. 309-314). Cambridge, MA: The MIT Press.

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

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.

Longo, G., & Montévil, M. (2014). *Perspectives on Organisms: Biological Time, Symmetries and Singularities*. Heidelberg, Germany: Springer.

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

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht: Kluwer Academic.

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.
Northoff, G. (2018). *The Spontaneous Brain: From the Mind-Body to the World-Brain Problem*. Cambridge, MA: The MIT Press.
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.
Roy, S., & Majumdar, S. (2022). *Noise and Randomness in Living System*. Singapore: Springer.
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.
Schurger, A., Hu, P. B., Pak, J., & Roskies, A. L. (2021). What Is the Readiness Potential? *Trends in Cognitive Sciences, 25*(7), 558-570.
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.