Mechanistic Explanations in the Cognitive Sciences: Beyond Linear Storytelling

Alexander Hölken

Abstract

Over the last two decades, an increasing number of cognitive scientists have turned to mechanistic explanatory frameworks in their efforts to describe and explain cognitive phenomena. Most mechanistic frameworks conceive of cognitive systems as composed of functionally-individuated components whose functions are narrowly defined by their ranges of possible inputs and outputs, as well as their relations to other components within the phenomenon-producing mechanism. In this paper, I argue that this *modular view* of cognitive mechanisms as linear systems is not applicable to biological cognitive systems, and offer an alternative characterization using the methodology of Dynamical Systems Theory.

Keywords: Dynamical Systems; Causation; Mechanisms

1. Introduction

Since the beginning of the 21st century, neo-mechanistic theories have steadily gained popularity within the cognitive sciences. These theories claim that the development and/or functioning of cognitive systems should be explained in terms of mechanisms and their components. For instance, a mechanistic explanation of how the visual cortex realizes object recognition would divide that cortex into a variety of sub-units such as cortical columns and retinotopic maps, and explain how each of their characteristic functions contributes to the recognition of a variety of visual objects.

Therefore, mechanistic explanations are successful when they correctly individuate the components that a mechanism consists of and give an accurate account of how it is realized by interactions between these components. As components may themselves be smaller-scale mechanisms (Craver, 2007), this means that mechanistic theories typically make reference to two different types of explanatory relations:

- *Etiological (same-scale) relations* between components or mechanisms whose output serves as input for another component or mechanism
- *Constitutive (inter-scale) relations* between smaller-scale components or mechanisms whose output *constitutes* part of a larger-scale mechanism's state or process

A simple example of an etiological relation would be that between the read/write head and the motor controlling the tape reels in a Turing machine: Based on the output from the read/write head (which results from its reading of the symbol on the tape underneath it), the tape reel motor spins either clockwise or anti-clockwise, moving the tape underneath the read/write head either left or right. This, in turn, presents the next symbol to the head. In this scenario, the read/write head and the motor are two same-scale systems interacting via one's output being the others' input.[1] An example of a constitutive relation would be the way in which the photoelectric sensor and the piece of graphite at the tip of the read/write head contribute to its ability to *function* as a device that can read and write symbols from a tape. In other words, this function is *constituted* by these two, smaller-scale components, standing in certain spatial and causal relations with each other. As such, constitutive relations obtain between components or mechanisms at different scales.

If these two types of causal relations were the only ones present in nature, mechanistic theories would provide excellent tools for explaining how any system comes to instantiate a certain function: We would be able to *neatly decompose* larger-scale mechanisms realizing more complex functions (e.g. visual feature recognition) into smaller-scale mechanisms realizing less complex ones (e.g. edge detection). This decomposition would be neat precisely when the components of a given mechanism can be *continuously individuated over time* and *robustly identified by their functional contribution*. However, while neat decomposition may be a successful strategy in the case of many artificial systems such as grandfather clocks, it fails to apply to more complex systems of the types found in chemistry and biology.

In this paper, I argue that mechanistic explanations, insofar as they refer to causal relations between physical systems producing a phenomenon of interest (POI), fail to explain the kinds of inter-scale causal processes evident in a wide range of natural systems. Starting with section 2, I discuss some of the basic requirements for mechanistic explanations and how they are rooted in the assumption that causal relations are *linear* in nature. In Section 3, I argue that this assumption of *linear causality* undermines their ability to provide causal explanations of what I term Complex Natural

[1] For the sake of brevity, this description is simplified. In actuality, the output of the head to the motor is mediated by the internal state register, and that of the motor to the head by the contents of the tape. Nonetheless, the explanation-relevant interactions are still between same-scale systems (the tape, the head, the register and the motor), there are just more of them.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society*.

Systems (CNS), followed in Section 4 by a comparison of linear and nonlinear causal-explanatory strategies in an example CNS. I close with a description of how nonlinear theories of causality avoid these issues and why they may serve as the foundation for establishing an updated version of *dynamicism* in the cognitive sciences, providing non-mechanistic explanations of cognitive phenomena.

2. Mechanistic Explanations and Linearity

Few example cases highlight the supposed *linear nature of causality* better than the classical philosophical example of the light switch, in which cause (flicking the switch) and effect (the light turning on) are connected by a literal (power) line. Using Fig. 1, which shows a simplified depiction of the linear relation between the flicking of the switch and the light's turning on, we can understand what a mechanistic explanation of this causal process could look like: If we wanted to explain why, for instance, flicking the switch causes the light to turn on, we could refer to the system consisting of the switch, the electrical circuit, the power source, and the light bulb, as a mechanism with two possible outputs (light on / light off). Which of these states occurs is then determined by two factors: The connections between the mechanism's components, called its *structure*, and the internal states of these components, or *subsystems*. Accordingly, manipulation of the two relations mentioned in Ch. 1 changes the light bulb's state: Flipping the switch up or down affects an *etiological relation*, since it serves as an input to the electrical wire, causing it to supply or cut off energy to the bulb. However, the state of the switch can only be causally efficacious with regards to the light bulb if the power source actually produces power, for instance by converting the energy from photons into electricity using the photovoltaic effect. Manipulating this process would affect the *constitutive relation* between the sub-components of the power source (a solar cell) and the functional contribution of that power source to the overall mechanism.

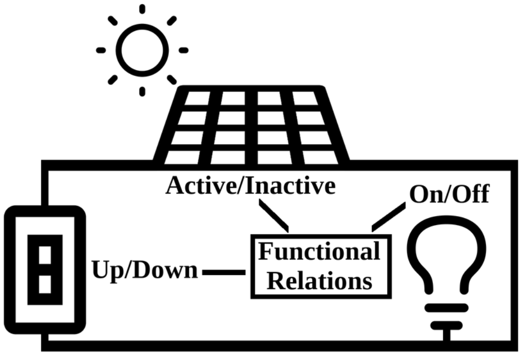

Fig. 1: Simplified example of a light circuit system. Mechanistic explanations of why the light turns on when the switch is flicked identify a set of causally-relevant subsystems based on their *functional contribution* to a causal chain ending with the light turning on. Thus, the state of the light does not depend on the subsystems' actual internal states, but on abstract relations between hypothesized functional states.

The light switch example showcases the minimum requirements for causal processes as they appear in mechanistic explanations, and linear causal explanations more generally. What makes a causal process *linear* is the fact that its development over time is determined completely by the two factors mentioned above: The structure of the system that realizes it (how the components of the electrical circuit are interconnected) and the states of the subsystems that constitute that system (the states of the switch, the bulb, and the power source). In other words, knowing the system's structure and the states of its subsystems, we can explain how etiological interactions between its subsystems and constitutive interactions between subsystem and whole-system states contribute to the light bulb's turning on.

Linear causal explanations thus depend on the assumption that subsystems act as *modules*: They causally contribute to realizing the larger-scale POI by producing certain outputs based on the input they receive and their current internal state – a process which can be identified with a *causal chain* of events (Fig. 2). Thus, when presented with a certain input, the subsystem in question will always develop in the same way, independently of how the other subsystems, or the system as a whole, develops.

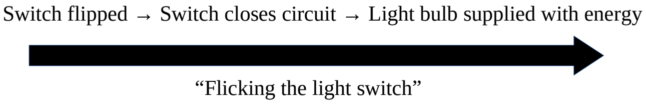
Switch flipped → Switch closes circuit → Light bulb supplied with energy
"Flicking the light switch"

Fig. 2: Example of a linear causal interpretation of the process of flicking the light switch. This process is composed of cause-effect couples that follow each other in linear time, which can be decomposed further to reveal how each sub-process contributes to constituting the overall causal chain.

Treating subsystems as modules has two practical upshots: First, it defines the etiological relations between subsystems not in terms of their *actual internal states*, but in terms of their *functional relations*: Sets of subsystems states are seen as interchangeable *insofar as they realize the same function* relative to other subsystems (i.e. mapping the same inputs to the same outputs). This decouples linear causal explanations from the concrete states of a given subsystem, instead giving explanatory weight to the *function* supposedly realized by that subsystem as part of a causal chain. In the case of mechanistic explanations, such functional relations even lead to the interpretation of different system developments in terms of normative aspects such as *system malfunctions* which can be traced down to *functional defects* in one or more subsystems. Second, the modular view affects how we experimentally manipulate (sub)system states in order to arrive at causal explanations: Since subsystem states are assumed to be independent of those of other subsystems (aside from the input states) and the system as a whole, we can explain a system's development sufficiently by *decomposing* it into its subsystems and manipulating their inputs individually to find out their range of possible states. Knowledge about each subsystem's range of states, coupled with knowledge about the way that subsystems are structurally related, is then sufficient for explaining any possible etiological and constitutive relation, if both of them are assumed to remain stable over time.[2]

Together, these analyses form the foundation for *linear causal explanations*: Such explanatory strategies first define a system realizing the phenomenon of interest (POI), which is then *neatly decomposed* into smaller-scale subsystems. Their contribution to causing the POI is then explained by identifying them with *functionally-defined modules within a causal chain* that may further be subdivided into simpler components with a more limited range of possible states. Explaining a given module's causal contribution to a POI is then simply a matter of locating it within a causal chain and, if needed, "zooming in" on its subsystems in order to see how their outputs result from the inputs they received, given their internal states and functional relations (Kellenberger, 2004). This explanatory strategy is linear in nature since it assumes that all subsystems causally interact on (roughly) the same spatiotemporal scale – in other words, it depends on the condition that actual causal processes occurring between subsystems can be understood as following each other in linear time, are not happening simultaneously, and that temporal dynamics (besides mere succession) play no causal role in their development (Port & Van Gelder, 1995).[3] As I argue in the following section, these assumptions are often unfounded in the case of CNS.

3. Limits of Linear Causal Explanations

While the assumption of linear causation may be useful for modeling purposes, or in cases where the development of a complex system needs to merely be approximately known, it is insufficient for a metaphysics of causation that can support causal explanations in Complex Natural Systems. To see why this is the case, we need to look at some of the metaphysical presuppositions of linear causal explanations. We can extract these from the example of the light switch:

- *Structural Fixation*: Within a system realizing a POI, both the ranges of possible states of its subsystems and the structural relations between subsystems remain fixed during the explanatorily relevant time period (den Hartigh et al., 2017).
- *Temporal Linearity*: An effect cannot causally influence its preceding cause: Causation always proceeds forwards in time along a causal chain, but never backwards (Kim, 1999).
- *Scale Specificity*: States and processes at different spatiotemporal scales do not causally interact with each other. Smaller-scale events, e.g. the solar cell's absorption of photons, do not cause, but constitute larger-scale ones (Glennan, 1996).

These three axioms of linearity jointly restrict the notion of causal processes to etiological relations only – that is, interactions between systems occurring in linear succession and (roughly) at the same spatiotemporal scale. Thus, most mechanistic explanations referring to causal interactions between components assume or imply the existence of linear causal relations between them. In addition, they conceive of constitutive relations as non-causal, completely eliminating the possibility of causal processes occurring between components at different scales (Krickel, 2018).

However, actual physical systems rarely fulfill all three axioms of linearity: For instance, scale specificity is a feature of linear systems only (as defined above), and thus not found in nonlinear systems which abound in nature (Bishop, 2008, 2012). Accordingly, linear causal processes understood in terms of etiological relations may, at best, act as abstractions or approximations of these nonlinear systems (Den Hartigh et al., 2017), on pain of failing to sufficiently explain – that is, allowing us to describe, manipulate and predict – POIs in Complex Natural Systems (Lamb, 2015).

For an illustration of the explanatory shortcomings of linear causal explanations, we can turn to biological systems, many of which do not display structural fixation and scale specificity. The Central Nervous System is a CNS in which nonlinear organizational features are particularly prominent: It frequently exhibits both structural adaptivity (from synaptic rewiring to changes in functional and effective connectivity) and a variety of electro-chemical processes simultaneously occurring at vastly different spatiotemporal scales, from those of a single neuron to whole-cortex oscillatory activity (Buzsáki, 2006). The structure of such systems does not remain fixed over time, as the relations between their subsystems constantly adapt to both changing states of other subsystems and of the larger-scale system they co-constitute. Rather than remaining static, *the structure of these systems is itself is a parameter for their development*, which renders linear causal explanations completely unsuited for explaining many of the phenomena they exhibit. In the following section, I refer to neural populations as a paradigm example of a CNS whose structure is a parameter for its development and argue for why we should abandon linear and mechanistic explanations in our effort to understand the complex, multi-scale causal relations that obtain in systems of this kind.

4. Neural Populations: A Case Study

The term *neural population* describes a group of neurons whose receptive fields correspond to sections along the same feature dimension. For instance, the activity of large groups of neurons in the deep layer of the *superior colliculus* corresponds to focal points within the visual field,

[2] Den Hartigh et al. (2017) call this the "control of variables" strategy. They argue that it does not succeed in explaining causal relations in anything but abstract models of natural systems, as both etiological and constitutive relations are highly adaptive and change over time within physically-instantiated natural systems.

[3] Of course, linear causal explanations may include the possibility for multiple different outcomes, e.g. via acyclic graphs (Glymour, 2001). However, such models are still linear in the mathematical sense, since they do not allow for the simultaneous realization of multiple branches or feedback between subsystem processes across different temporal scales.

with each neuron's tuning curve being "centered" around one two-dimensional coordinate specifying a point in the visual field. Together, these neurons realize a *retinotopic map* of focal targets, directing eye movements towards a particular point in visual space (Schöner & Spencer, 2016, p. 64f.). At first glance, this neural population may be considered an ideal candidate for a linear causal explanation: It receives a series of inputs from upstream areas and, based on the connectivity of its subsystems (neurons, interpreted as mechanisms), produces an output that is independent of any other system state. Taking this reasoning to its extremes, we could expect to be able to manipulate this population independently from other areas of the brain by providing it the correct kind of input, and reliably receive the same output (a focal target) in turn.

However, what we find in actuality is a different story: The neurons that constitute the entire population *do not act as functionally-individuated mechanisms*, with each of them playing the part of a module, situated within a linear causal chain. If this *were* the case, the causal contribution of each neuron to the overall mechanism of foveation would be to indicate something akin to *how closely its level of activation aligns with the location of the focal target* – that is, its causal role could be narrowly defined in functional terms. Accordingly, we would expect that a targeted chemical deactivation of such a neuron would make the population unable to direct the eyes to focus on that location in visual space. However, as shown by Lee et al. (1988), this is not the case – the deactivation of single neurons of the *superior colliculus* in monkeys did not impair foveation of the areas that these neurons were most attuned to (Fig. 3b). What's more, if a focal target was shown whose encoded location in the population was in proximity to the deactivated neurons, the resulting eye movements were biased *away* from the location coded for by the deactivated neurons – even though the neurons whose tuning curves corresponded most strongly to the focal target were completely intact (Fig. 3c).

What explains these findings is the fact that the location of the focal targets is not encoded by the activity of single neurons, but by activity patterns within the population *as a whole*: The activity of each neuron contributes to the creation of a *topological space* in which the region with the overall highest amount of activity defines the location of the focal target.[4] Therefore, an adequate causal explanation of how this population in the *superior colliculus* contributes to eye movements should not conceive of single neurons as modules with narrowly-defined functions (rendering them merely *functionally interchangeable*), but as *broadly interchangeable* elements of a larger-scale system which *realizes a topological space* within which the focal target locations are defined.

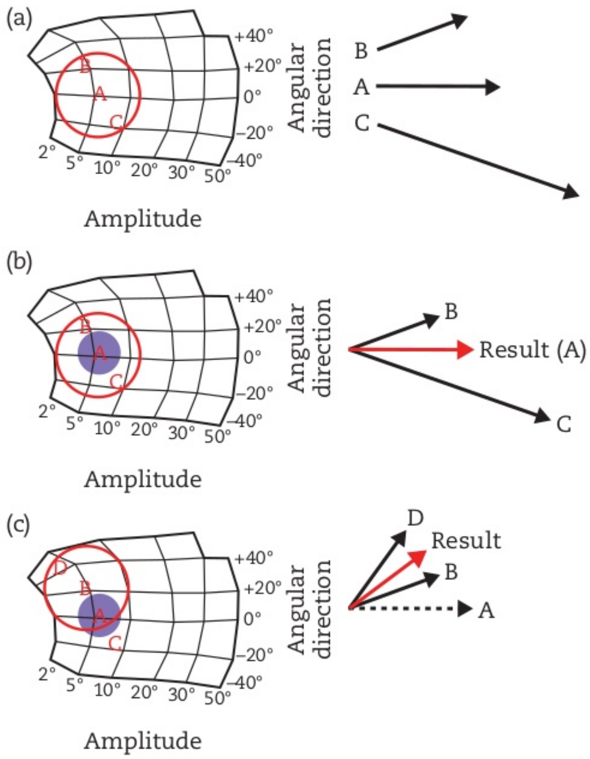

Fig. 3: Results of experiments by Lee at al. (1988), adapted by Schöner & Spencer (2016). On the left, the motor map realized by a population within the *superior colliculus* is shown, with the location of the deactivated neurons marked by a blue dot and the approximate region of neuronal activity elicited by the target stimulus shown by the red circle. (a) shows a condition in which no deactivation occurred, while in (b) and (c), neurons at location "A" were deactivated. On the right, the resulting eye movements are displayed as vectors, showcasing no impairment in condition (b) and a shift *away* from the deactivated area in (c).

Recall from section 2 that functional interchangeability is a direct result of the way in which mechanistic frameworks explain how mechanisms realize POI – namely by the characteristic functional contributions of the individual subsystems that make up the mechanism. Consequently, the function of the mechanism (bringing about the POI) is supposed to *causally result* from the functions of its subsystems. While this doesn't mean that the malfunctioning or deactivation of a *single* subsystem must necessarily cause the POI-realizing mechanism to fail, it does make mechanisms highly subsystem-dependent. In other words, if neural populations such as that within the *superior colliculus* were causal mechanisms, we would expect to be able to cause their functional breakdown with a small number of well-targeted interventions on functionally vital subsystems. However, what we see is a very different picture: Interventions such as the deactivation of groups of neurons or the repeated exposure to multisensory stimuli do not lead to a breakdown of function, but to *structural adaptation* (Yu et al., 2009). This is precisely a result of the fact that neural populations, like many other CNS, are not

[4] Dynamical Systems Theory refers to such regions as *attractors*, because they bias system development towards states within their vicinity. In neural systems, attractors and topological spaces are often transient, since they result from the *distribution of neural activity* throughout the population, not from the structural connections between neurons alone.

mechanisms defined by static structures and subsystems whose ranges of possible states remain fixed, but *self-organizing systems* whose subsystems jointly realize a dynamic structure whose topology results from the various constraints placed on it (Juarrero, 2023). The remainder of this paper clarifies this notion of a constraint-shaped topology and why it is often a part of successful explanations of phenomena in Complex Natural Systems.

5. Causal Explanations in CNS

As we have seen in the previous section, mechanistic approaches can, at least sometimes, be insufficient to explain a POI. In this section, I argue that this explanatory impotence of mechanistic explanations is the *rule*, not the exception, in the case of Complex Natural Systems. Instead of referring to *mechanisms* as explanatorily relevant entities, causal explanations in CNS should put *topologies* front and center, unless there are good reasons to look for alternatives.[5] Topologies, understood in this general sense, define biases in a *landscape of possible system states*, where states that the system is more likely to exhibit act as *attractors* (valleys) for nearby states, while states that the system is unlikely to exhibit act as *repellors* (hills).

Generally speaking, mechanistic explanations fail to explain a POI adequately if the system realizing that POI is not constituted by modular subsystems fulfilling a narrow range of functions, but by *broadly interchangeable* subsystems that self-organize into transient assemblies developing in a way that allows the overall system to adapt to external constraints (Sornette, 2006). For the purpose of this paper, we can understand these transient assemblies as groups of broadly interchangeable subsystems that entrain the development of other subsystems for as long as some external constraint needs to be satisfied (Anderson, 2015). Ecological niches serve as an accessible example of this process: Prey availability is an extremely relevant constraint for the behavior of a predator population within a specific ecological niche. If prey becomes rare, the spatio-temporal patterns of its occurrence act as a constraint on predator behavior – for instance, whether hunting occurs during the night or in packs. Since a lack of prey leads to individuals dying from starvation, the behavior of a subset of individuals – namely those that survive – gradually comes to dominate the behavior of the predator population as a whole. This situation persists until either the external constraint is relaxed (prey population increases) or another constraint becomes more dominant (e.g. changes in the local climate). As mentioned in section 3, Complex Natural Systems can also respond to a constraint by changing their internal structure or the range of states of their subsystems. In the case of a predator population, this might be realized by individuals adapting their behavior, e.g. by spending less energy on anything but hunting, or by the relations between individuals changing, e.g. a shift from individual hunting to pack hunting (Post et al., 1999; Fryxell et al., 2007).

The phenomenon of small groups of subsystems coming to dominate overall system development for a limited amount of time is a result of the ability of a CNS to exhibit *phase transitions* between different regimes (characteristic patterns of development). Despite phase transitions being ubiquitous in nature, mechanistic explanations cannot account for the concrete dynamics of these transitions within a physical system – at most, mechanisms can account for the multistability of such systems *in general*, by allowing that individual subsystems may realize multiple functions. However, questions about which precise conditions allow for the phase transition to occur, how it plays out and how it can be manipulated remain unanswered within mechanistic explanations that do not take the system's continuous temporal dynamics into account. For instance, investigating the predator population as a dynamical system, we can answer questions such as "Why did these wolves start pack hunting under these conditions?" with reference to the relevant causal topology of that system, which in this case is realized by the behavior patterns of individual wolves within the population.

Conversely, the predator system shows no evidence of a causal chain, constituted by functionally-individuated subsystems: No individual wolf realizes a narrowly-defined function as part of a causal chain realizing a mechanism, as in the light switch example. Rather, the development of the population as a whole *emerges* from the behavioral patterns of the individuals that constitute it, with the term *emergence* referring to the way in which subsystems realize a *causal topology* that is sensitive to a different range of states than its constitutive subsystems, and whose development causally affects these subsystems in turn (Haken, 1990). Thus, emergence can be defined as the ongoing process by which whole-system states, such as the hunting strategy of a wolf pack, are realized by, and in turn have an effect on, *relations between the states and processes of their subsystems*. These relations characterize precisely the structural adaptivity that linear causal explanations cannot account for. Nonlinear approaches, on the other hand, can explain how whole-system states (e.g. a hunting strategy) are both constituted by individual behaviors, and play a unique causal (and thus explanatory) role that individual behaviors do not.

The final nail in the coffin of linear causal explanations comes from the multitude of spatiotemporal scales at which CNS develop. In the case of the predator population, we saw that shifts in individual behaviors can cause changes on the level of the whole population, e.g. the hunting strategies. However, this causal process is *gradual*, both on the level of the individual and that of the entire population. It is

[5] Some empirical questions about POI in Complex Natural Systems may be best addressed by mechanistic explanations. However, these questions are often about *general features* of CNS, such as how neurons *in general* are able to generate spike trains based on their membrane potential. In contrast, mechanistic explanations often prove inadequate for explaining *concrete instantiations* of causal relations within such systems – e.g. why eliciting *this* membrane potential within a certain neuron led to *that* particular response pattern (cf. Beer, 2023).

important to note that the *rate at which these processes take place*, as well as their *spatial extension*, differs depending on which system instantiates them – the individual or the population. Generally speaking, systems extended over larger spatial scales (e.g. populations) develop more slowly than smaller-scale systems (e.g. individuals). This means that the development of systems and their constituent subsystems often occur *simultaneously, but on different spatiotemporal scales*. This results in the complete breakdown of the strict distinction between etiological and constitutive relations – it simply makes no sense to think of the relation between larger-scale CNS and their subsystems as one of asymmetric constitution when there is no single, universal spatiotemporal scale, relative to which changes to subsystem states may elicit changes in the larger-scale system's structure (Kiverstein & Kirchhoff, 2023). Instead, a system's causally-relevant states and processes occur *on a timescale whose resolution is relative to the temporal development of that very system.*

We can think of this temporal relativity in terms of a "perspective" of the system of interest: From the individual wolf's "perspective", the change of the population's hunting strategy doesn't occur as either a singular cause – a sudden change in how all other wolves hunt – nor a gradual, continuous development. Rather, the change in hunting strategy (a population-wide process) is causally efficacious on the level of the individual via a series of events that, from the perspective of the individual, seem unrelated or randomly patterned in time and space. Contrast this with how, from the "perspective" of the light bulb, changes of system-wide states or processes, such as the structure of the electrical circuit, occur either at singular points in time (such as when the circuit is interrupted) or as continuous changes over time (such as decreasing voltage due to wire corrosion). This is because most aspects of the light switch system can be adequately explained by reference to unique cause-effect relationships that constitute part of a causal chain between the different components of the system. Outputs of these components act as *efficient causes* on other components and can only lead to two kinds of results: Function or malfunction, defined in relation to the component's functional role within the mechanism. However, if a subsystem is not a functionally-individuated part of a mechanism, it is not constrained to a role of providing functionally-defined outputs based on inputs from other subsystems – in short, it cannot be seen as a modular component of a mechanism. Such subsystems do not fulfill narrowly-defined functional roles, acting as efficient causes like cogs in a man-made mechanism. Instead, they jointly realize a causal topology that *biases* the development of the whole system, which in turn affects their own development.

This kind of system-to-subsystem influence can be understood in terms of *teleodynamics* (Deacon & Koutroufinis, 2014) or *enabling constraints* (Juarrero, 1999), which affect the population's causal topology, thus biasing the range of possible states of each individual subsystem. Here, we see precisely the kind of variability of system structure and the subsystems' ranges of internal states that are central to nonlinear systems, exemplified by topological changes within the larger-scale system and the individual subsystems respectively. While such topology-based explanatory frameworks are still fairly novel (Kelso, 1995; Varela et al., 2001), they have seen increased application in both biology and the cognitive sciences since the dawn of the new century. A striking example of this framework in use can be found in DiFrisco & Jaeger (2020), who explicitly refer to topological regimes as causally efficacious entities that constrain the range of possible phenotypes realizable by a single gene.

Theoretical approaches such as these, which conceptualize topological features as causally efficacious factors for system development, have slowly but steadily gained popularity over the past few years. Similar to how *neo-mechanism* built on and reformed many of the elements of 20th-century mechanistic approaches to explanation and science, perhaps the theoretical and empirical research projects which make reference to causally efficacious topologies could be seen as heralds of a kind of *neo-dynamicism*, which aims to put into practice and improve on many of the mathematical and conceptual considerations made in the wake of then cutting-edge research on Dynamical Systems in the 1980s and 90s. While the initial hype around dynamical approaches has largely subsided, recent developments in various scientific fields concerned with living organisms have brought the issue of *causation* back to the forefront of many scientific research paradigms (Ladyman & Ross, 2009; Falkenburg & Morrison, 2015; Nicholson & Dupré, 2018). It is precisely this issue that dynamical accounts are furnished to address in a way that other approaches cannot – precisely because they are not limited to single-scale, linear explanations of causal relations. An in-depth engagement with the unique type of causality present in Complex Natural Systems will be required to discern to which degree topological explanations may replace widely-accepted mechanistic ones in such systems, and whether the two approaches may, in fact, be commensurable (see for instance Huneman, 2018).

6. Conclusion

In this paper, I argued that mechanistic explanations of empirical POI necessarily make reference to functionally-individuated components (subsystems) whose interaction forms part of a causal chain. I then argued that this picture is incomplete at best when it comes to explaining phenomena in Complex Natural Systems, whose subsystems are not functionally-individuated, but rather self-organize into assemblies that maintain, and are in turn affected by, a system-wide topology that serves to ensure system coherence in the face of a varied set of constraints. Finally, I argued that this topology should be understood as a causally efficacious entity in its own right, and that such a conceptualization could serve as the basis for a neo-dynamicist approach to explanation that can provide a realistic alternative to currently-popular mechanistic ones.

References

Anderson, M. L. (2015). Beyond Componential Constitution in the Brain: Starburst Amacrine Cells and Enabling Constraints. *Open MIND*. https://doi.org/10.15502/9783958570429

Beer, R. D. (2023). On the Proper Treatment of Dynamics in Cognitive Science. T*opics in Cognitive Science*, tops.12686. https://doi.org/10.1111/tops.12686

Bishop, R. C. (2008). Downward causation in fluid convection. *Synthese*, 160(2), 229–248. https://doi.org/10.1007/s11229-006-9112-2

Bishop, R. C. (2012). Fluid convection, constraint and causation. *Interface Focus*, 2(1), 4–12. https://doi.org/10.1098/rsfs.2011.0065

Buzsáki, G. (2006). *Rhythms of the brain*. Oxford University Press.

Craver, C. F. (2007). *Explaining the Brain*. Oxford University Press.

Deacon, T., & Koutroufinis, S. (2014). Complexity and Dynamical Depth. *Information*, 5(3), 404–423. https://doi.org/10.3390/info5030404

Den Hartigh, R. J. R., Cox, R. F. A., & Van Geert, P. L. C. (2017). Complex versus Complicated Models of Cognition. In L. Magnani & T. Bertolotti (Eds.), *Springer Handbook of Model-Based Science*. Springer International Publishing. https://doi.org/10.1007/978-3-319-30526-4

DiFrisco, J., & Jaeger, J. (2020). Genetic Causation in Complex Regulatory Systems: An Integrative Dynamic Perspective. *BioEssays*, 42(6), 1900226. https://doi.org/10.1002/bies.201900226

Falkenburg, B., & Morrison, M. (Eds.). (2015). *Why More Is Different: Philosophical Issues in Condensed Matter Physics and Complex Systems*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-43911-1

Fryxell, J. M., Mosser, A., Sinclair, A. R. E., & Packer, C. (2007). Group formation stabilizes predator–prey dynamics. *Nature*, 449(7165), 1041–1043. https://doi.org/10.1038/nature06177

Glennan, S. S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, 44(1). https://doi.org/10.1007/BF00172853

Glymour, C. (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. MIT Press.

Haken, H. (1990). Synergetics as a Tool for the Conceptualization and Mathematization of Cognition and Behaviour—How Far Can We Go? In H. Haken & M. Stadler (Eds.), *Synergetics of Cognition* (Vol. 45, pp. 2–31). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-48779-8_1

Huneman, P. (2018). Diversifying the picture of explanations in biological sciences: ways of combining topology with mechanisms. *Synthese,* 195, 115–146. https://doi.org/10.1007/s11229-015-0808-z

Juarrero, A. (1999). *Dynamics in action: Intentional behavior as a complex system*. MIT Press.

Juarrero, A. (2023). *Context changes everything: How constraints create coherence*. The MIT Press.

Kellenberger, E. (2004). The evolution of molecular biology: Biology's various affairs with holism and reductionism, and their contribution to understanding life at the molecular level. *EMBO Reports*, 5(6), 546–549. https://doi.org/10.1038/sj.embor.7400180

Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. MIT Press.

Kim, J. (1999). Making Sense of Emergence. *Philosophical Studies*, 95, 3–36. https://doi.org/10.1023/A:1004563122154

Kiverstein, J., & Kirchhoff, M. (2023). Dissolving the causal-constitution fallacy: Diachronic constitution and the metaphysics of extended cognition. In *Situated cognition research: Methodological foundations* (pp. 155–173). Springer International Publishing.

Krickel, B. (2018). *The Mechanical World: The Metaphysical Commitments of the New Mechanistic Approach* (Vol. 13). Springer International Publishing. https://doi.org/10.1007/978-3-030-03629-4

Ladyman, J., & Ross, D. (2009). E*very thing must go: Metaphysics naturalized.* Oxford university press.

Lamb, M. (2015). *Characteristics of Non-reductive Explanations in Complex Dynamical Systems Research.* University of Cincinnati.

Lee, C., Rohrer, W. H., & Sparks, D. L. (1988). Population coding of saccadic eye movements by neurons in the superior colliculus. *Nature*, 332(6162), 357–360. https://doi.org/10.1038/332357a0

Nicholson, D. J., & Dupré, J. (Eds.). (2018). *Everything flows: Towards a processual philosophy of biology* (First edition). Oxford University Press.

Paul, L. A. (2007). Constitutive Overdetermination. In J. K. Campbell, M. O'Rourke, & H. S. Silverstein (Eds.), *Causation and Explanation*. The MIT Press. https://doi.org/10.7551/mitpress/1753.003.0015

Port, R. F., & Van Gelder, T. (Eds.). (1995). *Mind as motion: Explorations in the dynamics of cognition*. MIT Press.

Post, E., Peterson, R. O., Stenseth, N. Chr., & McLaren, B. E. (1999). Ecosystem consequences of wolf behavioural response to climate. *Nature*, 401(6756), 905–907. https://doi.org/10.1038/44814

Schöner, G., & Spencer, J. P. (2016). *Dynamic thinking: A primer on dynamic field theory*. Oxford University Press.

Sornette, D. (2006). *Critical Phenomena in Natural Sciences*. Springer-Verlag. https://doi.org/10.1007/3-540-33182-4

Varela, F., Lachaux, J.-P., Rodriguez, E., & Martinerie, J. (2001). The brainweb: Phase synchronization and large-scale integration. *Nature Reviews Neuroscience*, 2(4), 229–239. https://doi.org/10.1038/35067550

Yu, L., Stein, B. E., & Rowland, B. A. (2009). Adult Plasticity in Multisensory Neurons: Short-Term Experience-Dependent Changes in the Superior Colliculus. *The Journal of Neuroscience*, 29(50), 15910–15922. https://doi.org/10.1523/JNEUROSCI.4041-09.2009

Acknowledgements

Fig. 3 reproduced with permission from Schöner & Spencer (2016). Many thanks to Maja Griem, Julia Wolf, and Abdullahi Ali for their feedback on early versions of this paper, as well as three anonymous reviewers for their valuable contributions.

This work was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) – Project No. GRK-2185/1 (DFG-Graduiertenkolleg Situated Cognition).