Explicating Top-Down Causation Using Networks and Dynamics

William Bechtel
Department of Philosophy and Center for Circadian Biology
University of California, San Diego

Abstract

In many fields in the life sciences investigators refer to *downward* or *top-down* causal effects. Craver and I defended the view that such cases should be understood in terms of a constitution relation between levels in a mechanism and intra-level causal relations (occurring at any level). We did not, however, specify when entities constitute a higher-level mechanism. In this paper I appeal to graph-theoretic representations of networks, now widely employed in systems biology and neuroscience, and associate mechanisms with modules that exhibit high clustering. As a result of such interconnections mechanisms often exhibit complex dynamic behaviors that constrain how individual components respond to external inputs, a central feature of top-down causation.

Keywords: constraints; downward causation; endogenous dynamics; graph representations; mechanistic explanations; negative feedback

Contact Information: William Bechtel, Department of Philosophy, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0119; bechtel@ucsd.edu

Acknowledgements

I thank three anonymous referees for this journal for their very helpful comments and suggestions. I also thank John Norton and Visiting Fellows at the Center for Philosophy of Science at the University of Pittsburgh in 2014-2015, especially Sara Green, Raphael Scholl, and Maria Serban, for their spirited discussion of an earlier draft of this paper. Likewise, I thank members of the audience at the Workshop on Levels of Organization, Causality, and Top-Down Relations sponsored by the IAS Research Center for Life, Mind, and Society, University at the Basque Country, San Sebastian, in June 2015, especially Leonardo Bich, Alvaro Moreno, and Kepa Ruiz-Mirazo, for valuable discussion at and after the Workshop. Finally, I thank Jason Winning for many productive discussions concerning constraints and mechanism.

1. Introduction

States of whole systems often constrain the behavior of their parts. Conditions in a cell, such as its phase in the cell cycle, constrain which genes are expressed. The same molecule can have different effects on components of a cell (e.g., promoting or inhibiting apoptosis) depending on the conditions in the cell. Phenomena such as these are often characterized as involving *downward* or *top-down causation* (Noble 2006). They are contrasted with cases of bottom-up causation in which a state of a component of a system partially determines the state of the whole (e.g., a genetic mutation impairs fatty-acid metabolism of a cell or bonds between actin and myosin produce muscle contraction). While most researchers do not find bottom-up causation mysterious (it is invoked in reductionistic explanation of properties of a system in terms of its parts), many have found top-down causation to present a puzzle: how does a whole system have effects over and above those of each of its components?

In an attempt to defuse the mystery, Craver and I (Craver and Bechtel 2007) distinguished constitution and causation and proposed treating constitution as a relation between lower-level parts and higher-level mechanisms and causation as a relation between entities at the same level. On our proposal, parts constitute wholes; they don't cause their properties. By viewing causation as occurring at all levels, not just the lowest one, we presented themselves as offering an account that would capture what is characterized as top-down causation without engendering conceptual problems such as those posed by Kim's (1998) exclusion argument. When one thing acts on a whole mechanism, and its components are also modified, we proposed that higher-level causation was involved in producing the effect on the whole mechanism. Since the parts are what constitute the whole mechanism, one or more of them would themselves be changed in that process. No additional causal processes were involved between the whole and the part, although additional causal processes might then ensue in the mechanism as a result (thereby altering the state of the system as a whole).

However, the examples Craver and I introduced fail to bring out in what sense higher-levels are involved in producing these effects at the lower level. In offering an example in which a person's activity (playing tennis) results in a change within the person's body (altered metabolism), we emphasized the role of lower-level causal relations: "In this and many similar cases, a change in the activity of the mechanism as a whole just is a change in one or more components of the mechanism which then, through ordinary intra-level causation, causes changes in other components of the mechanism." This presents the challenge: why should one treat the whole mechanism as a higher-level entity rather than just a collection of lower-level entities, with all the causality operative between these lower-level entities? This is a version of Kim's exclusion argument against top-down causation and a number of critics (see, for example, Soom 2012; Rosenberg 2015) has objected that in the end Craver and me, like Kim, only allow causation at the lowest level. In pushing a similar objection, Fazekas and Kertész (2011) argue that constitution should be expressed in an identity claim, and this undermines any sense of autonomy for higher-level causation.

My goal in this paper is to unpack Craver and my distinction between causation and constitution by explicating more clearly (1) when an entity or activity that is regarded as at a higher-level enters into causal relations such that the causality should be treated as at a higher-level and (2) the relation that holds between the state of the mechanism as a whole and the state of its components. While bottom-up causation has seemed less mysterious, the same problem arises in cases of supposed bottom-up causation: how do operations of parts of a system have effects on the whole when what they seem to have is effects on other parts, which together constitute whole. Whereas top-down causation raises the question why the whole is considered the cause, bottom-up causation raises the question why effects are assigned to the whole.

Crucial to any account of inter-level causation is the notion of level that is invoked. Craver and I dissociated our treatment of levels from many in the literature, such as the notion of levels of science that were invoked in the theory-reduction literature or notions of levels defined in terms of the size of entities—e.g., molecules, cells, organs, organisms (for an frequently cited example, see Churchland and Sejnowski 1992). In the context of discussions of top-down causation, neither of these senses brings out what many find to be problematic. There is nothing problematic with objects studied in one discipline having causal effects on those studied in other disciplines or for large entities to have causal effects on small entities. Craver and I restrict our account to mechanistic levels in which a mechanism consists of appropriately organized parts performing operations. The notion of a mechanism causing the state of its parts does bring out what many have found problematic in talk of top-down causation since the state of the parts and the state of the mechanism are not independent in the sense required for one to cause the other.

What is crucial to the mechanistic conception of level is the idea of an entity being constituted by its parts and operations. The notion of constitution is what must be explicated and section 2 will highlight this challenge by developing an example biological mechanism that I will use through the rest of the paper. Then, in section 3, I introduce a framework for thinking about systems that has been extensively applied in systems biology (Barabasi and Oltvai 2004) and neuroscience (Sporns 2010): graph-theoretic representations of networks. Appealing to graph theory may appear paradoxical as graph-theoretic representations do not explicitly advert to levels—all nodes are represented on a plane. Many graphs that characterize biological systems, though, involve modules in which there is high clustering of nodes often around one or more hubs. Many of these modules, on the account I am offering, constitute higher-level mechanisms. One must not only show that high-level mechanisms can be identified, but that the effects on the parts when the whole mechanism is affected correspond to what has led to talk of top-down causation. That is, the condition of the whole mechanism must result in different behavior of the part than would occur when the conditions in the whole mechanism are different. This requires that we turn from the structure of modules to their functioning. The clustered nodes in biological networks are typically not ordered sequentially; rather, they are connected in such an interconnected manner that one can identify multiple feedback loops. When the operations corresponding to the edges in graphs are non-linear, interconnected modules

can exhibit complex dynamic behavior such as oscillations. As a result of dynamic behavior within the module, the module itself does not respond to external inputs in the same manner on all occasions. How it responds depends on the current conditions of the module. This is a diagnostic symptom of top-down causation.

The graph representations that I introduce in sections 3 and 4 help make it clear when it is appropriate to identify mechanisms as higher-level components—mechanisms are highly interactive modules within a larger network that are capable of exhibiting complex dynamics. But as a result of treating all interactions, those within the module and those between modules, as edges, the graph representation does not make manifest why the effects of mechanisms on their constituent components are different from the propagation of causal effects throughout the network. Instead, the graph representation may reinforce the perception that all causation is at the lowest level represented by individual nodes in the graph. To address this issue, in section 5 I will deploy a distinction between constraints and dynamical laws that has been introduced into theoretical biology from physics (Hooker 2013; Pattee 1971). Constraints reduce the degrees of freedom that are left open by dynamical laws alone by, for example, establishing correlations between variables or restricting the range of variables. Although the imposition of constraints and the propagation of effects of constrained systems involve diachronic activity, the constraints exercised by a whole mechanism on its parts are synchronic—they are realized through the organizational relationships between the components at a time. Given the organization of the mechanism, the responses of components that are altered when the mechanism itself is affected by external causes are constrained by the whole.

2. When Top-Down Causation Seems Problematic: Levels of Mechanisms

The notion of mechanism invoked in the conception of levels I am addressing arises from the practice of biologists over the past several centuries. In the life sciences, investigators developing explanations often (1) begin by identifying the mechanism responsible for a specific phenomenon to be explained, (2) proceed to decompose the mechanism into its parts and the operations they perform, and (3) finally recompose the mechanism to show how, as a result of the organized parts orchestrating their operations, the mechanism generates the phenomenon. While this practice has been pursued for several centuries, it has assumed a central place in philosophical accounts of explanation in the last couple decades (Bechtel and Richardson 1993/2010; Bechtel 2006, 2008; Machamer, Darden, and Craver 2000). The steps of decomposition and recomposition are what invite employing the word *levels*. Parts are, in a rather natural sense, at a lower level than the mechanism constructed out of them.

This conception of level can be illustrated using an example to which I will return throughout this paper, the phenomenon of circadian rhythmicity in mammals (i.e., daily oscillations in behaviors and physiological functions that are endogenously generated but entrainable to day-night cycles in the local environment). The responsible mechanism resides in individual cells. The genes *Per* and *Cry* and the proteins synthesized from them,

PER and CRY, are major parts.[1] PER and CRY form a dimer and, after being transported back into the nucleus, inhibit their own transcription by interfering with the activators, BMAL1 and CLOCK.[2] These genes and proteins occupy a lower level than the mechanism itself.

Using this example, we can now see what is sometimes regarded as problematic about top-down causation. Whether *Per* and *Cry* are transcribed and translated into the proteins PER and CRY depends upon the phase of oscillation the host cell is in. The oscillatory phase, a state of the whole cell, is determining the behavior of its parts, seemingly in accord with accounts of top-down causation. However, the phase of the oscillator at a time just is the concentrations of PER, CRY, and a cadre of related proteins that make up the mechanism. Treating the concentrations as causing the phase, or vice versa, seems to violate many intuitive aspects of causation. Causation is often understood as involving contact action or a propagated signal (Hitchcock 2003). Moreover, causes are assumed to precede their effects. Both conditions require that causes and effects be wholly distinct (see Lewis 2000, , for detailed arguments as to why causes and effects must be distinct). The phase of the oscillation in the cell is not distinct from the concentrations of PER and CRY. More generally, the parts and wholes of a mechanism are not distinct—each requires the existence of the other. There is no possibility for transmission between parts and wholes since they are not distinct. Neither the states of the parts nor the state of the whole come before the other—they occur simultaneously. Talking of top-down causation in these cases also seems to engender the problem of redundant causation highlighted by Kim (1998). While we might attribute the change in concentration in PER or CRY to the phase of the oscillator, it can also be attributed to specific molecular events, *viz.*, that the proteins have attached themselves to two other proteins, BMAL and CLOCK, and removed those proteins from the E-boxes on the *Per* and *Cry* genes, terminating the synthesis of more PER and CRY.

Craver and I sought to make a virtue out of these problems for top-down (or comparable problems for bottom-up) causation by treating the inter-level relation as constitution and noting that a state of the whole mechanism involves at least some of its parts being in appropriate states. Any phenomenon that seems to exhibit bottom-up or top-down causation can be accommodated by viewing causes as operating either (1) between the whole mechanism and other entities outside it or (2) among the parts within the mechanism. The fact that the parts constitute the whole has the consequence that the state of the whole mechanism is changed whenever the state of one of its parts is changed and the state of at least one of the parts is changed whenever the state of the whole is changed. Craver and Bechtel referred to how the constitution relation mediates between changes to the parts and changes to the whole as yielding *mechanistically mediated effects*.

---

[1] There are two paralogs of PER and CRY in the mammalian clock mechanism, but since they function in similar ways, I am omitting this detail.

[2] There is now some doubt as to whether the mechanism requires proteins inhibiting their own transcription; on some proposals, the oscillator may be post-translational. For purposes of this paper, I will assume that the core mechanism involves transcription-translation feedback loops.

A major shortcoming in Craver and my account is that by merely pointing to a constitution relation we left unspecified what it is for parts to constitute a mechanism such that the mechanism is at a higher level. The need to address this question can be seen by examining the three-tier diagram Craver (2007) uses to present his view of levels. At the top an arrow on the left terminates at a darkened oval that represents a mechanism. Another arrow leaves from the mechanism. This is intended to show that when the mechanism receives an input, it generates an output. Dotted lines connect that oval to the one below, indicating that what is below is an expansion of what is above. In this middle tier, the arrow on the left is shown as entering into the mechanism to terminate at one of the ovals within it. The four ovals within are connected by arrows, culminating in one from which the arrow on the right now exits. One of the ovals in the mechanism is again shown in black, and it is expanded in the third tier in the same manner. The figure makes clear the sense in which this account of mechanisms is supposed to give rise to a view of levels—the inner ovals in tier two are inside the larger oval which is then shown in black above it on the upper tier. The question the diagram poses is: what do the ovals represent? Put another way, why is the oval in the middle tier around all four inner ovals, and not some other possible combination of ovals (e.g., just two of the inner ovals in tier 2, or one of these and another outside of the oval shown)?

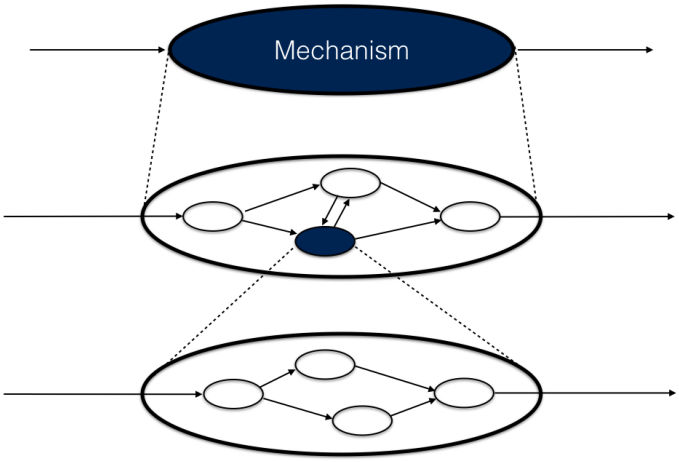

Figure 1. Craver-style representation of three levels of a mechanism. Adapted from Craver (2007).

Motivating the ovals in Figure 1 requires an answer to the original question: when do components constitute a mechanism? A mechanism is not just an aggregation of parts, each of which performs an operation. In their characterization of mechanisms, Machamer, Darden and Craver (2000) emphasize productive continuity between the parts involved in generating the phenomenon. This is critical, but not sufficient. A pond is a collection of entities, many performing operations, with productive continuity between operations. But it is not what scientists would call a mechanism. One thing that is central in all accounts of a mechanism is that the parts are those entities whose activities or operations are responsible for the phenomenon attributed to the whole mechanism. In the mechanism literature, the specification of the phenomenon to be explained is often invoked to determine which entities belong to a mechanism and which, even though they are located among the other entities, do not. With respect to the pond, if one identifies a phenomenon,

such as maintenance of pH, then one can search for the mechanism—the productively linked operations of entities in the pond that contribute to that phenomenon.

As useful as appealing to a phenomenon to be explained is in identifying candidate parts of a mechanism, it is insufficient to fix the boundaries and hence pick out a mechanism as a higher-level entity existing as such in the world. One reason is that the same entity may be involved in the production of several different phenomena. To accommodate this, we require an account in which a mechanism can share parts with other mechanisms. Another reason is that the range of entities or activities that can affect a given phenomenon is not sharply bounded. Often entities very distant in time and space can play critical roles in the generation of a phenomenon (Bechtel 2015). The first entity in tier two of Craver's diagram has an arrow coming into it from somewhere else, but that entity is, for some unexplained reason, not counted as part of the mechanism.

The issue of where to draw boundaries around a mechanism is in fact a crucial issue in biology. As research proceeded on circadian rhythms researchers identified multiple feedback loops in addition to that involving PER and CRY. For example, BMAL1, which serves as an activator of *Per* and *Cry* transcription, is produced by a feedback loop in which its synthesis is regulated by the nuclear receptors RORα and REV-ERBα, while it together with CLOCK are activators of their synthesis. Figure 2 presents the conception of the core clock mechanism as it was understood around 2005. It presents the circadian clock mechanism as a well-bounded set of components. (The figure itself does not show any inputs or outputs, but in fact there are input signals that serve to entrain the clock to the day/night cycle and output signals that serve to regulate the expression of a wide variety of other genes.) Since then the emergence of new techniques, such as knocking down expression of genes through use of small interfering RNAs, has revealed more than 300 additional genes that are both expressed in a circadian manner and exert effects on the phase or amplitude of circadian rhythms (Zhang et al. 2009). Many of these genes were already identified as components of other cellular mechanisms. The last dozen years has revealed extensive interactions, for example, between components of basic metabolic mechanisms and core components of the circadian clock. The questions of where to draw the boundary of the clock mechanism and why to draw it there have assumed prominence (for examples, see Bechtel 2015). If Craver's account is to provide a principled characterization of entities in terms of mechanistic levels, one needs a procedure for limning the boundaries of the mechanism and distinguishing it from other entities that affect the phenomenon but are not identified as parts of the mechanism.

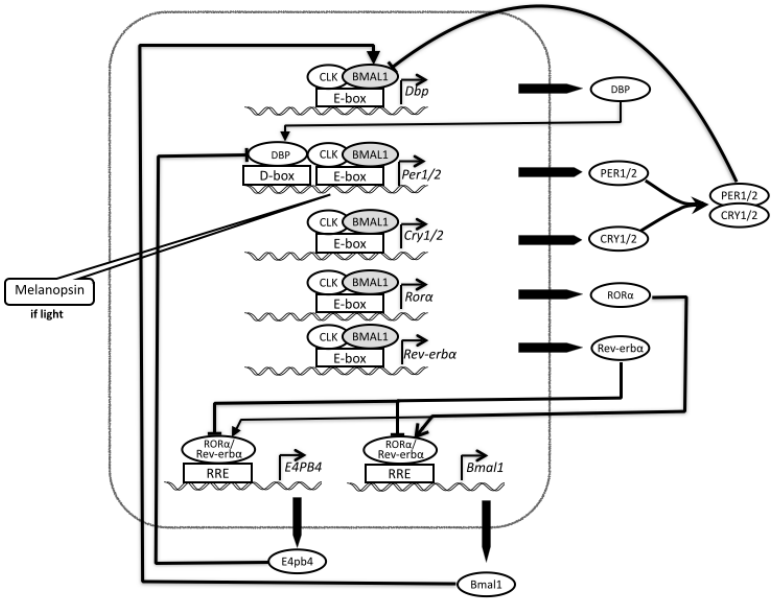

Figure 2. The major parts and operations of the mammalian circadian clock as understood circa 2005. Not shown are the various kinases that phosphorylate the proteins and determine whether they are broken down or transported into the nucleus.

3. Flattening Levels: Using Graphs to Identify Mechanisms

I will return to the question of what the ovals represent in Craver's diagram below. First, I turn to a second question posed by his diagram. At the top level, arrows just contact the black oval representing the mechanism. At the middle tier, the corresponding arrows contact the smaller ovals inside the larger oval. At the lowest tier, the arrows penetrate the larger oval that corresponds to one of those small ovals and contacts yet inner ovals. What the penetration of arrows into the mechanism at lower tiers suggests is that the causal effect is not on the higher-level unit as a whole, but on one (or possibly several) of its components. If in the bottom tier, instead of expanding an inner oval, Craver had chosen to expand the initial oval from the middle tier, the point would have been even clearer. Then on the three tiers the arrow representing input to the mechanism would connect, respectively, with the highest-level oval (the mechanism), one of the inner ovals (a component of the mechanism), and one of the ovals within that oval (a component of the component). At the lowest tier, the arrows in and out are no different than those between components and the idea of causation at multiple levels seems to be lost. In many cases, such reduction to lower levels might seem to be appropriate—when light exposure entrains circadian rhythms, photons affect the melanopsin molecules in the intrinsically photoreceptive retinal ganglion cells. As a result, when neurotransmitters are released they initiate a signaling cascade within cells in the suprachiasmatic nucleus, resulting in increased *Per* transcription. All of these events seem to be at a single level. But then in what

sense is the clock, a higher-level entity, entrained? And how does its phase as a result of entrainment, affect the behavior of its components?

This exegesis of Craver's diagram suggests that the critics who viewed Craver and my account as rendering higher levels epiphenomenal were right. It suggests a highly reductionistic picture of levels according to which causal relations that were supposed to be between entities at higher levels of organization dissolve into causal interactions at the lowest level considered. To bring out this point, I have presented a flattened representation of Craver's diagram in Figure 3. It preserves the relationships between components at the middle and lowest tier in Craver's diagram, and fills in possible decompositions of the three ovals in the middle tier that were not expanded in Craver's diagram. To identify what were supposed to be higher-levels entities, I have included dotted ovals around the lower-level ovals and a dashed oval around the whole set.

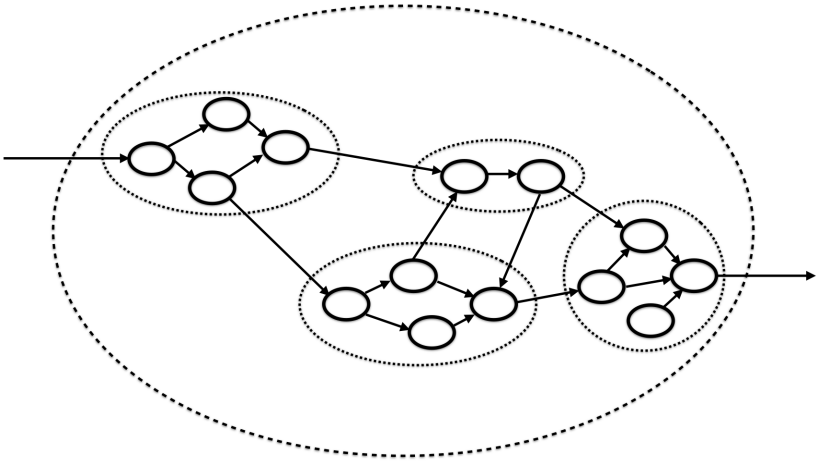

Figure 3. A flattened redrawing and elaboration of Craver's diagram, with dotted ovals grouping nodes that correspond to units at the middle and top level in Craver's diagram.

If one ignores the dotted ovals, Figure 3 corresponds to a graph-theoretic representation of a network. A graph representation uses nodes (ovals) to represent entities and edges to represent relations (perhaps causal) between nodes. In constructing a graph representation, one has to settle on which entities are to be represented as nodes. Often in systems biology the nodes represent relatively low-level entities such as genes or proteins. Here I have simply taken the entities in the bottom level in Craver's diagram as nodes. This does not presuppose that there is a lowest level (or, as I will discuss below, that all entities are at the same level). As inquiry proceeds, researchers may elect to treat the entity represented as a node as a mechanism and replace the node with a set of nodes constituting its constituents. Edges may represent either structural connections or functional connections (for purposes of this paper I will assume they are functional). In graph representations more generally, the edges in a graph may be directed or undirected, but in a graph representation of a mechanism, the edges are directed and represent operations in which one node exercises a causal effect on another.

By showing all edges as between nodes, the graph representation brings out what is challenging in explicating top-down or bottom-up causation—there does not seem to be

any principled criterion for identifying levels. The dotted and dashed ovals that correspond to entities at higher levels in Craver's diagram appear to be purely arbitrary impositions on the graph representation. One could draw ovals that group nodes in different ways. Moreover, since they are not the endpoint of edges, the dotted and dashed ovals or whatever they represent seem to be causally inert. Rather than providing an account of causation at multiple levels, it appears that all higher levels appear to have been rendered epiphenomenal.

I will argue, however, that there is a way to use graph-theoretic analysis to identify structures within networks that correspond to the sort of entities traditionally viewed as residing at higher levels and to show in what respect these entities constrain causal processes. To do so, I need to introduce three measures graph theorists have introduced to characterize graphs:

1. *Mean shortest path-length* is a measure of the average number of edges that must be traversed on the shortest path between two nodes.
2. The *clustering coefficient* is a measure of how connected to each other the nodes linked to a given node are.
3. *Degree distribution* is a measure of how the number of connections from different nodes is varied.

In terms of these measures, one can describe different network topologies. A randomly connected network will have a small mean shortest path-length (there are, on average, short paths between any two units), allowing for rapid transmission between nodes, but exhibit little clustering of units. Lattice or near-neighbor structures exhibit high clustering, allowing nodes to combine operations to produce collective effects, but have long mean shortest path-lengths. Highly clustered units are often referred to as *modules*. Networks that Watts and Strogratz (1998) designated *small worlds* retain the short mean shortest path-length of random networks but contain modules of clustered units more typical of near-neighbor networks. Figure 4 presents a toy example network in which nodes represent entities and arrows causal interactions between them. It exhibits small-world organization—there are distinct modules of interconnected units but a number of connections linking components in different modules. As discussed further below, the interconnections within these modules enables them to function as higher-level units in which the response to an input from outside depends on the state of the module.

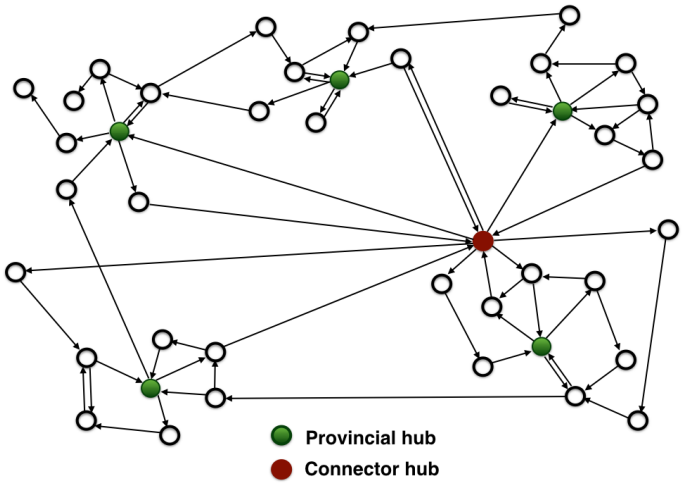

Figure 4. A graph representation of modules organized around provincial hubs and integrated through a connector hub.

When these network topologies were being explored, most researchers assumed that node degree would be distributed normally. However, Barabási and Albert (1999) found that in many real world networks, degree distribution corresponds to or approximates a power law: most units have few connections to other units, but some units are very highly connected to other units. Networks with power-law degree distribution are referred to as *scale-free* as there is no characteristic scale on which to describe the network. Highly connected nodes are commonly referred to as *hubs*. *Provincial hubs* are ones that exhibit high clustering and so serve as integrators of activity in modules. *Connector hubs*, on the other hand, often have low clustering and serve to integrate activity between modules. Although the number of nodes is too small to produce a truly scale-free network, the graph in Figure 4 illustrates how modules can be organized around provincial hubs and integrated together through a connector hub.

What is important for present purposes is that the modules in these networks are differentiated on principled grounds—the nodes within a module are more interconnected with each other than they are with nodes elsewhere. The extensive interactions between the nodes constituting a module enable them to work together as a unit. This does not mean that the connections to components outside the module are not important; inputs from them may figure critically in regulating some behaviors of the module and enabling the output of the mechanism to affect other processes. But the enhanced connectivity enables greater coordination within the module, allowing the components to work together and justifies treating them as constituting a mechanism when they produce a phenomenon we seek to explain. To return to Zhang et al.'s (2009) discovery of an additional 300 genes that affect circadian rhythms when knocked down, it is noteworthy that they did not view them as core clock components since they are not nearly as interconnected as the components shown in Figure 2. Rather, they viewed the proteins expressed by these genes as external factors that interact with the mechanism. One of the interesting consequences of the application of network analyses in systems biology is that often researchers identify

clusters or modules within them that correspond more or less closely to mechanisms that have been identified using more traditional approaches in cell and molecular biology. Ravasz et al. (2002) provide one example of this. The researchers constructed an overlap matrix for the metabolic network in *E. coli* and identified substrates that form clusters or modules based on edges corresponding to metabolic interactions. The modules identified through this analysis closely approximated those traditionally identified as constituting mechanisms, but often included components not previously identified.

4. From Graphs to Dynamics: Coordinated Dynamics in Mechanisms

So far I have tried to show that biological mechanisms more closely correspond to interconnected modules in scale-free small-world networks (Figure 4) than to ovals in Craver's diagram (Figures 1 and 3). To appreciate the importance of interconnected modules, we must move beyond graph-theoretic representations and consider the types of dynamic behavior such organization supports. When one starts with one part and traces connections in interconnected networks, such as those shown in Figures 2 and 4, one often discovers that the parts affected by the operation themselves perform operations that directly or indirectly affect the operation of the part from which one started. That is, the operations feed back onto parts that were envisioned as earlier in the process. Such feedback does not involve backwards causation, since the effects are not on current but future operations of the part, but over time it can give rise to coordinated dynamical behavior that constrains the behavior of the parts of the module.

Understanding feedback has proven challenging for humans. Although we know it was already employed by Ktesibios of Alexandria in the 3rd century CBE to regulate the flow of water in his water clock, negative feedback did not become recognized as a general design principle until the 20th century, when cyberneticists such as Wiener (1948; Rosenblueth, Wiener, and Bigelow 1943) presented it as a principle for enabling engineered and natural systems to achieve target outcomes. Prior to that it had been rediscovered many times, including by Watt, who employed it in his governor for the steam engine, which inspired to Maxwell's mathematical analysis of governors (Mayr 1970). But even when negative feedback was recognized as a design principle for regulating systems to pursue target outcomes, many theorists did not attend to the fact already recognized by engineers that negative feedback could generate oscillations.

Oscillations in biological organisms were often concealed by such techniques as examining mean behavior and not attending to time-series. Physiological processes were envisioned, in accord with Machamer, Darden, and Craver (2000), as proceeding "from start or set-up to finish or termination conditions." Variation in activity was regarded as noise. But increasingly through the 20th century researchers came to recognize that a broad range of physiological processes, from glycolysis to neural processing, generate oscillations. Following on models, such as one advanced by Goodwin (1965), negative feedback was recognized as a design principle for generating endogenous oscillations and when oscillations were discovered in biological organisms, investigators proposed negative feedback mechanisms. Circadian rhythms were no exception, and circadian researchers

were on the hunt for a feedback mechanism for several decades before Hardin, Hall, and Rosbash (1990) provided critical empirical evidence.[3] Using cloning, they demonstrated that the mRNA and proteins produced by the first identified circadian gene, *Period* (*Per*) oscillated with a circadian period, with the concentration of the protein peaking several hours after the mRNA. On this basis, they proposed the feedback loop described above in which PER inhibits the transcription of *Per*. Mental simulation of the operations proposed reveals how the mechanism could oscillate: when PER levels are low, synthesis of new PER proceeds, but as PER accumulates, it inhibits further synthesis. Only when sufficient PER has broken down can synthesis resume. Mental simulation cannot determine whether the oscillations will dampen or be sustained indefinitely (assuming a sufficient supply of free energy); accordingly, Goldbeter (1995) constructed a computational model, inspired by Goodwin's, that demonstrated that the proposed negative-feedback mechanism could produce sustained oscillations under physiological conditions.

What is important for thinking about top-down causation is that the overall state of the circadian mechanism determines how components within it behave and how they will respond to perturbations arising outside the mechanism. This state of the whole mechanism might be described in terms of the states of some components of the mechanism (for example, PER concentrations in the nucleus are high). The effect on *Per* transcription at that time is determined by conditions generated within the mechanism, not from outside it. This point, however, extends far beyond the circadian example. Any network in which the edges are not all in one direction is subject to complex dynamical behavior, such as oscillation, in which the behavior of parts is constrained by the behavior of other parts in the network.

The constraining effect of the whole on the parts is clearly seen in how the parts respond to external inputs differentially depending upon the state of the mechanism. This is manifest in the process by which the circadian clock is entrained to local conditions, especially light conditions. Light has different effects at different times of day, as exhibited in the phase-response curve shown in Figure 5 (time is shown in Circadian Time, according to which 0 corresponds to dawn). During the early part of the night a light pulse delays the phase of the oscillation, but a light pulse late at night advances the phase. During daytime light pulses have no effect. Once details of the clock mechanism in mammals were worked out, researchers determined that entrainment to light resulted when a light signal from the retina to the suprachiasmatic nucleus functions to increase *Per* transcription. Now researchers could understand why light exposure had different effects on the phase depending on time of day. If *Per* transcription is already at its maximum, as it would be during anticipated daytime, light input could have no further effect. At dusk and early night, as *Per* concentration is diminishing due to the feedback mechanism, a light signal can counter the increasing inhibition and keep transcription going longer. This has the effect of delaying the phase of the oscillator. On the other hand, later in the night as dawn is approaching, *Per* concentration is again beginning to increase due to the endogenous

[3] They used fruit flies as their model organism, but within the decade the homologs of *Per* were found in mammals.

oscillation. A light signal at this phase will speed up transcription, resulting in reaching daytime levels in PER concentration earlier and advancing the phase of the oscillator.

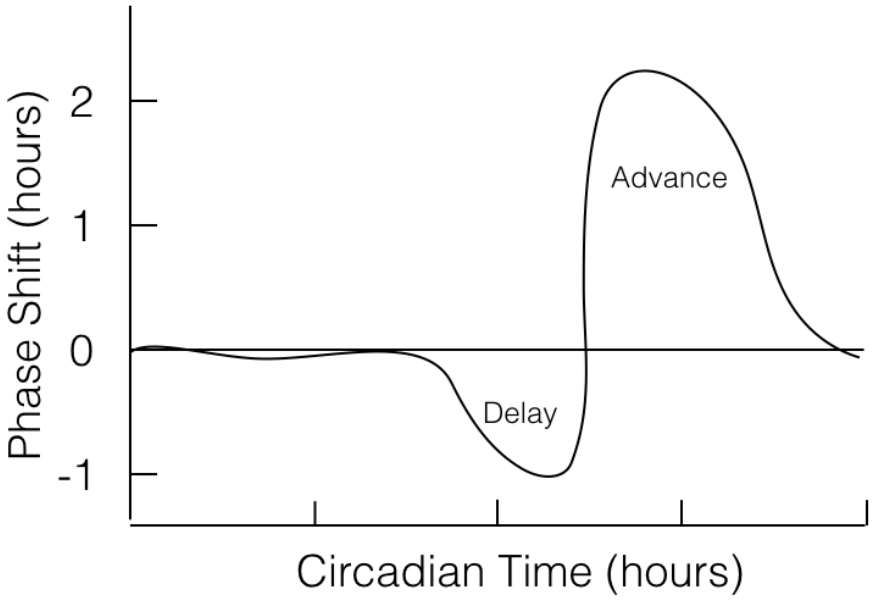

Figure 5. Typical phase response curve for rodents kept for total darkness and then administered light pulses at different times. Circadian time 0 is the time of expected dawn.

The circadian example makes clear the functional importance of the type of integration of nodes found in modules in network representations of biological systems. The operations of individual components (e.g., transcribing genes) depend on the state of other parts of the mechanism (e.g., the concentrations of the PER protein in the nucleus). As a result of the interconnectivity of the parts, especially the feedback loops, the module identified as the mechanism functions as a unit, with the operations of the individual parts of the mechanism determined by other parts of the mechanism. If there were no feedback loops, the parts that receive the input would not be sensitive to conditions elsewhere and would always respond in the same manner to the input. There would also be little reason to identify the parts as constituting a module. The feedback loops are responsible for the state of the module modulating the responsiveness of its parts to external input. Even though the input to the mechanism may only affect one or a few components, the mechanism as a whole is the relevant unit due to the interconnections that run through the mechanism.

As network analyses have been pursued in fields such as systems biology and neuroscience, it has become apparent that modules often form a hierarchy: modules often have far more connections to some modules than others. The toy network in Figure 4 shows interconnections among all the modules that render the whole into a module. The interconnections that give rise to these larger modules also result in dynamics that affect the behavior of the individual modules. A last circadian example provides an illustration of such a hierarchy and how it contributes to control over individual components multiple levels below. An important feature of oscillatory systems, already recognized by Huygens in

the case of pendulum clocks, is that a very weak signal is capable of causing the oscillations to synchronize. I noted above that the suprachiasmatic nucleus (SCN), a bilateral structure with about 10,000 neurons on each side, serves as the master circadian clock in mammals. By dispersing SCN neurons, Welsh et al. (1995) determined that individual neurons sustain oscillations but with widely varying periods. When they are connected in the SCN or in slices, they synchronize with each other, thereby enabling sloppy oscillators to generate a highly regular signal to transmit to the rest of the organism (Aton and Herzog 2005). Synchronization of components depends on the long-distance connections exhibited in Figure 4. It begins with operations in which molecules such as the hormone VIP are synthesized in individual cells. These are dispersed out of the cell to bind to receptors on other cells, initiating a signaling cascade in those cells that ultimately affects the processes of *Per* transcription and translation within these cells.

To understand synchronization, one must identify modules at two different levels, and understand how conditions in modules at each level affect both their own components and those at still lower levels. The interactions between SCN neurons determines the rate and efficiency of synchronization among them, while the interactions within neurons determines how they respond when VIP binds to one of their receptors. Although the individual operations are described at the molecular level, they are constrained by the current conditions in the large-scale modules.

5. Top-Down Effects Due to Constraints Imposed by Networks Dynamics

My strategy in the previous sections has been to use network analyses to identify modules and the dynamic behavior that arises in some modules with rich interactive connections to characterize the constitutive relation that holds between parts and a mechanism. When network analyses are applied to biological systems, one frequently finds small-world organization in which node degree is distributed according to a power law. This has the result of creating modules of highly interconnected nodes in which, nonetheless, several nodes still receive inputs from outside the module. These modules often correspond to biological mechanisms that have been identified through more classical techniques.

In Craver's diagram (Figures 1 and 3), the ovals drawn around mechanisms seemed arbitrary, but on the network account there are principled reasons for picking these out as modules. What renders a group of entities into a module is the interconnections and interactions between them. When the entities work together to produce a phenomenon, they count as a mechanism. The interconnections and interactions often yield dynamic behavior in which components in a mechanism behave differently at different times due to activity elsewhere in the mechanism. As a result, determining the organization and dynamics of the module is crucial for understanding the behavior of its parts when they receive external inputs.

A concern I raised earlier about Craver and my treatment of top-down causation is that it rendered all causal relations at the lowest level. To first appearances, graph theoretic representations of networks seem to reinforce that concern. In cases in which things

outside the mechanism exercise effects on the mechanism, it will typically be by affecting one or more parts of the mechanism. These altered parts then causally modify other components of the mechanism. Even endogenous activities such as oscillations are explained in terms of the feedback between components of the mechanism. The network representation seems to favor a highly reductionistic account that represents all activity at one lowest level.

This interpretation, however, is mistaken. First, the nodes in a network need not belong to a common level in any of the standard senses. In some cases a graph representation of a network is developed by identifying interactions between entities that might be taken to be at a common level on grounds such as common size or common type of entity. Gene regulation and protein interaction networks are examples. In other cases the nodes correspond to entities that are structurally or functionally connected independently of whether they are situated at what is regarded as a common level. The only sense of level that is explicitly embodied in a network diagram is between nodes and modules that comprise them. This is the familiar mechanistic sense of level. As Craver and I argued, however, this provides only a very local conception of level—the parts of a mechanism that interact are at the same level. This account provides no guidance for judging whether the nodes of different modules correspond to entities at the same level or whether the sub-parts of the parts belong to the same level. On this conception of level, there are no grounds for treating the nodes in a given graph as at a common level.

Second, although in any graph representation there will be a set of nodes that correspond to what are taken as the basic entities, they should not be treated as representing entities at some base level. At best they represent the entities at which the graph representation bottoms out. Just as researchers investigating a mechanism have a choice as to whether to characterize the behavior of the mechanism as it interacts with other entities, many of them mechanisms (thereby elaborating the characterization of the phenomenon) or to decompose it and appeal to its parts and operations to explain its behavior, those developing a graph representation have a choice as to whether to represent a whole mechanism as a node, or decompose it into other nodes representing the parts of the mechanism. On many occasions researchers seek to decompose one part of the mechanism, leaving others untouched. The graph representation will show the components into which the one mechanism has been decomposed interacting with the other mechanisms that have not been decomposed.

Third, in developing a graph representation, one might deliberately represent a set of entities as a single node. Researcher may have already identified the components of a mechanism, but deem them not to be relevant to their analysis. For example, in analyzing the interactions between SCN neurons, researchers may choose to treat the neurons as units, ignoring the interactions of molecules within them. The graph representation will employ units for neurons and edges for the connections between them. There are contexts in which the whole SCN becomes a single node and the edges are the connections between the SCN and the various organs that the SCN regulates and which, in many cases, send inputs back to the SCN. The graph representation format does not privilege a lowest level,

but represents as nodes those entities whose interactions are deemed relevant for understanding the phenomenon of interest.

Although a graph representation provides a convenient way to identify modules in networks, by using the same type of arrow to relate nodes within a module and those between nodes in different modules, it glosses over an important distinction relevant to understanding top-down causation in mechanisms. Looking diachronically both the edges between nodes inside a module and those between modules may represent causal relations. But in the moment when the mechanism receives causal input from outside by having the state of one or more of its parts altered, the relation between the parts and the mechanism as a whole is not diachronic but synchronic. At a given time, the mechanism is constituted of its various parts. The natural language to use to talk about how synchronically the parts are affected by the whole is that of *constraint*—being situated in the mechanism constrains the behavior of the part.

The notion of constraint has roots in mechanics. Fundamental laws such as those proposed by Newton characterize possible ways a system might evolve given initial conditions. But in a given situation constraints restrict those possibilities, foreclosing some while leaving others open (Hooker 2013). Constraints such as an inclined plane limit the motion a marble can take as a result of the gravitational attraction between it and the earth. In terms of mathematical representations of the evolution of a system, constraints are captured in relations added to dynamical equations that limit the degrees of freedom available through which the system might evolve.

Some constraints, such as an inclined plane or an electric wire, are fixed with respect to the system in question. They provide a minimalist notion of top-down causation. The path a marble takes after being deposited on an inclined plane is governed by the angle of the plane. Other constraints change, often in response to other activities in the system. A switch in an electrical system can direct electricity along different paths in a circuit. More interesting sets of constraints, and more interesting examples of top-down causation, arise as parts of a system constrain each other in ways that change dynamically. In the much-discussed example of Bénard cells, after the application of heat, molecules begin to move and exert constraints on each other. Eventually coordination between the molecules results in a macro-scale convection pattern in which individual molecules are constrained.

The notion of constraint has been applied to biology by a number of authors (Pattee 1971; Rosen 1985; Hooker 2013; Moreno and Mossio 2015). A key foundation of their thinking is that biological organisms exist far from thermodynamic equilibrium and in order to build and maintain themselves, they must constrain the flow of free energy to perform work that builds the structures that perform the constraint role. The different mechanisms constituting a biological organism each play a role in such a process by restricting the range of activities that can occur. The circadian mechanism that provided my example throughout this paper figures in such a network of constraints as it constrains various physiological processes (typically by regulating expression of other genes) to occur at times when they can work together to maintain the organism. The parts of the mechanism itself are

generated by the operation of the mechanism (the protein PER is synthesized when BMAL1 binds to the promoter on the *Per* gene) and perform operations in it (PER inhibits the ability of BMAL1 to activate the transcription of the *Per* gene).

The notion of constraint provides a way to understand how the constitution of a mechanism results in the phenomena referred to as top-down causation. Appealing to graph theoretic representations of systems, I have identified mechanisms as dynamical systems that can arise in modules. The degrees of freedom available to entities of such interactive dynamical systems are reduced and so they behave differently than they would if not part of the system. Especially when the constraints are changing as a result of the dynamical activity, these entities may exhibit different behavior on different occasions.

Thinking in terms of constraints also facilitates a response to the assumption of the closedness of the physical that lies at the foundation of Kim's exclusion argument. Fundamental dynamical laws do apply universally and as long as one can specify initial conditions for each entity in the system, one can determine its behavior by invoking these laws. But even to address problems of analytic mechanics, physicists recognize the need to invoke constraints. These constraints are not derived from the laws themselves but must be ascertained empirically and added to the laws to determine behavior. The set of possible constraints is not closed. Yet, only in terms of constraints can one predict or explain outcomes. Moreover, they are determined locally. In biology, the constraints imposed in a mechanism are specific to the conditions in the living system. From this perspective, the physical is far from closed but rather is extremely open-ended. Wherever one finds a set of components organized into a module with sufficient interactions, one will encounter constraints that limit the behavior of the components and how they respond to external inputs. The phenomenon described as top-down causation is not unusual, but common.

6. Conclusion

Craver and I proposed that bottom-up as well as top-down causation could be understood by limiting causal processes to within levels of mechanisms and treating the constitution of mechanisms as mediating between levels. We did not, however, provide an account of what it was about constitution of mechanisms that motivates appeals to effects or causes at other levels. Invoking graph representations, in this paper I have characterized mechanisms as modules that appear in many graphs of biological networks. Modules consist of nodes that are highly clustered and in that respect are distinguished from other nodes (with which they still have a number of edges). Particularly important in integrating nodes into modules that can exhibit what are taken to be top-down effects are feedback connections that are common in biological systems. With feedback, modules are capable of endogenous activity that results in them being in different states at different times. When this happens, some of their components are in different states. Depending on the state, the component will produce different responses to inputs. This is the sort of behavior that has characteristically motivated talk of top-down causation. In developing this account, I have employed flat graph representations of systems in which modules can be identified. Their dynamics can then be analyzed. This has allowed me to pick out mechanisms in which

phenomena traditionally characterized as top-down causation can be understood without directly appealing to levels. But this explication of levels also allows us, if desired, to reintroduce reference to levels and characterize what is meant by top-down or bottom-up causation in non-problematic ways. Finally, by characterizing how a mechanism affects parts in terms of constraint I have differentiated how the mechanism restricts the behavior of its parts from the way the parts are affected by external inputs and provided a way to resist the assumption that the closedness of the physical underlying Kim's exclusion argument.

References

Aton, Sara J., and Erik D. Herzog. (2005). "Come Together, Right...Now: Synchronization of Rhythms in a Mammalian Circadian Clock." *Neuron* 48:531-534.

Barabási, Albert-László, and Réka Albert. (1999). "Emergence of Scaling in Random Networks." *Science* 286:509-512.

Barabasi, Albert-Laszlo, and Zoltan N. Oltvai. (2004). "Network Biology: Understanding the Cell's Functional Organization." *Nature Reviews Genetics* 5:101-113.

Bechtel, William. (2006). *Discovering Cell Mechanisms: The Creation of Modern Cell Biology*, Cambridge: Cambridge University Press.

———. (2008). *Mental Mechanisms. Philosophical Perspectives on Cognitive Neuroscience*, London: Routledge.

———. (2015). "Can Mechanistic Explanation Be Reconciled with Scale-Free Constitution and Dynamics?" *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*.

Bechtel, William, and Robert C. Richardson. (1993/2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*, Cambridge, MA: MIT Press. 1993 edition published by Princeton University Press.

Churchland, Patricia S., and Terrence J. Sejnowski. (1992). *The Computational Brain*, Cambridge, MA: MIT Press.

Craver, Carl F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*, New York: Oxford University Press.

Craver, Carl F., and William Bechtel. (2007). "Top-Down Causation without Top-Down Causes." *Biology and Philosophy* 22:547-563.

Fazekas, Peter, and Gergely Kertész. (2011). "Causation at Different Levels: Tracking the Commitments of Mechanistic Explanations." *Biology & Philosophy* 26:365-383.

Goldbeter, Albert. (1995). "A Model for Circadian Oscillations in the *Drosophila* Period Protein (Per)." *Proceedings of the Royal Society of London. B: Biological Sciences* 261:319-324.

Goodwin, Brian C. (1965). "Oscillatory Behavior in Enzymatic Control Processes." *Advances in Enzyme Regulation* 3:425-428.

Hardin, Paul E., Jeffrey C. Hall, and Michael Rosbash. (1990). "Feedback of the *Drosophila Period* Gene Product on Circadian Cycling of Its Messenger Rna Levels." *Nature* 343:536-540.

Hitchcock, Christopher. (2003). "Of Humean Bondage." *British Journal for the Philosophy of Science* 54:1-25.

Hooker, Clifford A. (2013). "On the Import of Constraints in Complex Dynamical Systems." *Foundations of Science* 18:757-780.

Kim, Jaegwon. (1998). *Mind in a Physical World*, Cambridge, MA: MIT Press.

Lewis, David. (2000). "Causation as Influence." *Journal of Philosophy* 97:182-197.

Machamer, Peter, Lindley Darden, and Carl F. Craver. (2000). "Thinking About Mechanisms." *Philosophy of Science* 67:1-25.

Mayr, Otto. (1970). *The Origins of Feedback Control*, Cambridge, MA: MIT Press.

Moreno, Alvaro, and Matteo Mossio. (2015). *Biological Autonomy: A Philosophical and Theoretical Inquiry*, Dordrecht: Springer.

Noble, Denis. (2006). *The Music of Life: Biology Beyond the Genome*, Oxford: Oxford University Press.

Pattee, Howard Hunt. (1971). "Physical Theories of Biological Co-Ordination." *Quartery Review of Biophysics* 4:255-276.

Ravasz, Erzsébet, A. L. Somera, D. A. Mongru, Zoltan N. Oltvai, and Albert-Laszlo Barabasi. (2002). "Hierarchical Organization of Modularity in Metabolic Networks." *Science* 297:1551-1555.

Rosen, Robert. (1985). "Organisms as Causal Systems Which Are Not Mechanisms: An Essay into the Nature of Compleixty," In Robert Rosen, ed., *Theoretical Biology and Complexity: Three Essays on the Natural Philosophy of Complex Systems*, 165-203. New York: Academic Press.

Rosenberg, Alex. (2015). "Making Mechanism Interesting." *Synthese*:1-23.

Rosenblueth, Arturo, Norbert Wiener, and Julian Bigelow. (1943). "Behavior, Purpose, and Teleology." *Philosophy of Science* 10:18-24.

Soom, Patrice. (2012). "Mechanisms, Determination and the Metaphysics of Neuroscience." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43:655-664.

Sporns, Olaf. (2010). *Networks of the Brain*, Cambridge, MA: MIT Press.

Watts, Duncan, and Steven Strogratz. (1998). "Collective Dynamics of Small Worlds." *Nature* 393:440-442.

Welsh, David K., Diomedes E. Logothetis, Markus Meister, and Steven M. Reppert. (1995). "Individual Neurons Dissociated from Rat Suprachiasmatic Nucleus Express Independently Phased Circadian Firing Rhythms." *Neuron* 14:697-706.

Wiener, Norbert. (1948). *Cybernetics: Or, Control and Communication in the Animal and the Machine*, New York: Wiley.

Zhang, Eric E., Andrew C. Liu, Tsuyoshi Hirota, Loren J. Miraglia, Genevieve Welch, Pagkapol Y. Pongsawakul, Xianzhong Liu, Ann Atwood, Jon W. Huss, Jeff Janes, Andrew I. Su, John B. Hogenesch, and Steve A. Kay. (2009). "A Genome-Wide Rnai Screen for Modifiers of the Circadian Clock in Human Cells." *Cell* 139:199-210.