4-23-2018

Constraints and Explanation

Alexander Bolano
alexbolano92@gmail.com

CONSTRAINTS AND EXPLANATION

by

Alex Bolaño
University of Missouri-St. Louis
B.A. Philosophy – Saint Louis University

A Thesis submitted to the Graduate School of the University of Missouri-St. Louis in partial fulfillment of the requirements for the degree
Master of Arts in Philosophy

May, 2018

Advisory Committee

Lauren Olin, Ph.D
Chairperson

Jon McGinnis, Ph.D

Gualtiero Piccinini, Ph.D

Since the deductive-nomological (D-N) model to scientific explanation fell out of favor, causal and mechanical approaches to scientific explanation have undergone a surge in popularity. For example, Salmon (1984) writes, “Causal processes, causal interactions, and causal laws provide the mechanisms by which the world works; to understand why certain things happen, we need to see how they are produced by these mechanisms” (p. 132). Similarly, Elster (1983) argues that “causal explanation is the unique mode of explanation in physics” (p. 18). Contemporary philosophical accounts of scientific explanation such as Woodward (2003), Craver (2007), and Strevens (2008) emphasize the importance of describing causal phenomena for scientific explanations.

Recently though, there has been much philosophical work done on the topic of non-causal explanation in science. Philosophers such as Baker (2005), Pexton (2015), Lange (2016), and Chirimuuta (2017) all offer persuasive examples of genuine non-causal explanations in the sciences. In this paper, I want to examine a special class of non-causal explanations in the sciences, what I have dubbed *constraining explanations.* Constraining explanations work by showing how the explanandum follows as a direct consequence of formal/mathematical constraints on the kinds of objects, processes, or interactions that can populate a model. A common example of this kind of explanatory pattern involves appeals to “invariance principles” in physics. Invariance principles, such as symmetry principles and conservation laws, are kinds of “super principles” that coordinate and frame individual natural laws (Wigner 1972 p. 996). Invariance principles work by constraining the possible mathematical forms that natural laws and phenomena can take.

The layout of this paper is as follows. In section 1, I give a brief overview of the main issues surrounding philosophical accounts of scientific explanation. In section 2, I present some examples of scientific explanations and argue that they work primarily by constraining the possible causes, entities or states of affairs that we can expect to find in the world. In section 3, I argue that these explanations give us a kind of modal knowledge that causal explanation alone cannot give us. Constraining explanations show us how a particular explanandum would have still occurred, even given a number of possible different causal antecedents. Constraining explanations give the explanandum phenomenon a kind of stability and invariance that causal-mechanical explanation alone cannot give.

1. A Brief History of Explanation – Laws and Causes

The natural place to start on a philosophical discussion of explanation is the famous (infamous?) deductive nomological (D-N) model of explanation given by Hempel & Oppenheim (1948). The basic gist of the D-N model can be summed up in one sentence: scientific explanations are like deductive arguments. According to the model, an explanandum *E* is explained if and only if *E* follows as a logical consequence of at least one law of nature *L1, L2...Ln,* in conjunction with antecedent background conditions *C1, C2...Cn.*

Here is a very simple explanation to illustrate this model. Say we are trying to explain why this particular sample of gas expanded when we put a flame under it. In this case, the explanandum *E* is "This sample of gas expanded". The relevant law *L* in this situation is the law "All gasses expand when heated" and the relevant antecedent

conditions *C* is “this sample of gas was heated.” *E* is logically entailed and can be predicted on the basis of *L* and *C*, and so according to the D-N model, we have explained *E*.

*O*ne major criticism of the D-N model of explanation is that it cannot account for the apparent asymmetrical direction of explanation. Consider this example, let us say we are trying to explain the length *l* of a shadow cast by a flagpole. To do this, we would appeal to the height *h* of the flagpole, and the relevant law *L* that represents the relationship between the height of the object, the angle of light, and the length of the shadow. So far so good. Notice, however, we can substitute *l* for *h* in the preceding argument, and logically derive *h*. For the D-N model, an explanation is sufficient only if the explanandum is logically entailed by the explanans. So, according to the D-N model, we have just explained the height of the flagpole in terms of its shadow.

Obviously, it does not seem to make much sense to explain the height of the flagpole in terms of the length of the shadow it casts. The explanatory relation seems only to work one way; from the height of the flagpole to the length of the shadow, not vice versa. Therefore, something must be missing from the D-N model. Many philosophers have suggested the missing ingredient is *causation*. The height of the flagpole explains the length of the shadow because the height of the flagpole (in conjunction with the light from the sun) *causes* the shadow to be a particular length. The perceived importance of causation in regard to explanation has led some philosophers such as Salmon (1994), Woodward (2003), Craver (2006) and Strevens (2008), to advance theories of scientific explanation that place causation center stage.

Of course, philosophers have different ideas on what exactly constitutes a causal explanation. Some philosophers treat causal explanations as proceeding by identifying the constitutive entities and mechanisms that underlie the phenomenon (Craver 2006 & 2009). Others focus on describing causal processes (Salmon 1994), and other focus on accounting for the objective dependence relationships that obtain between the elements of causal systems (Woodward (2003) & Strevens (2008)). Causal mechanical approaches to explanation seek to explain the world like one would explain the activity of a mechanical clock; try to figure out how all the pieces fit and move together. In general, causal approaches to modeling hold that the more accurately a given model represents the causally relevant features of the target system, the more explanatory the model is. Kaplan (2011), for example, introduces what he calls the *model-to-mechanism-mapping* (3M) account for modeling. According to the 3M account, a model is explanatory of some phenomena if there is a mapping between elements of the model and elements of the mechanism for the phenomenon (p 340). In other words, models should aim to represent the constitutive mechanisms of a system and the causal relations that obtain between those mechanisms.

The impetus for causal-mechanical approaches to explanation is well motivated. Many sciences, especially the special sciences such as biology and neuroscience, frequently and successfully explain in terms of causes, mechanisms, or processes. However, I want to argue that causal-mechanical approaches to scientific explanation cannot be the whole story. There are explanations in science that do NOT derive their explanatory power from identifying causes, mechanisms, or processes. Instead, these

explanations work, roughly, by identifying constraints on what kinds of causes, mechanisms, or processes that there can be. These explanations work by identifying formal principles of scientific models that directly constrain the possible kinds of objects, properties, events, or laws that can be included that model. A formal principle, as I understand the term, is a general principle, often characterized in mathematical terms, that expresses a dependency relation between elements of a system. Formal principles often set limitations on the set of mathematical structures we can use to model a system. In a sense, these explanations not only tell us about the way that the world actually is, but also about the ways that the world *could not be*. Constraining explanations can do this because they delineate the space of possible explanandum that we can expect to find in the world. The two major classes of examples of constraining explanations I discuss concern the role that invariance principles play in modern physics, and the role that optimization models play in evolutionary biology. Both examples that I consider involve explaining broad general features of a class of phenomena by appealing to formal mathematical principles of models.

2. Constraining Explanations: Invariants and Optimalities

2.1 Invariance Principles in Physics

Consider this example, given by Lange (2013): We are asked to explain why mother fails every time she tries to divide 23 strawberries evenly among her three children. The explanation is the fact that 3 does not evenly divide 23 (p. 488). The mathematical fact “3 does not evenly divide 23” makes it impossible that mother could ever succeed in her task of dividing the 23 strawberries evenly among her 3 children. The

mathematical fact “3 does not evenly divide 23” constrains the possible outcomes that could happen in this situation. Specifically, the mathematical fact constrains the possible outcomes to precisely only those outcomes where she fails to equally distribute the 23 strawberries.

Let’s turn this strawberry situation into a very simple mathematical model. The equation $s/x = n$ models the relationship between the total amount of strawberries s, the number of children n, and the strawberry dividing technique x, where it is understood that s, x, and n are whole numbers. It is easy to see that no x can satisfy this equation when $s = 23$ and $n = 3$. The model, in virtue of its formal characteristics alone, rules out situations where mother succeeds in dividing the 23 strawberries evenly into 3 groups. Consequently, this mathematical fact explains why mother fails every time she tries to distribute the strawberries in the manner specified. The model does not “sanction” such a state of affairs as legitimate.

This constraining explanation shows how the explanandum phenomenon follows directly from mathematical constraints on the particular form that a particular phenomena can take. The formal relationship between the number of strawberries present and the number of children present tells us something about the possible states of affairs we can expect to see in the world. This relationship tells us that we can never expect to see a situation in which mother succeeds in dividing the strawberries evenly. Conversely, the relationship also tells us that every possible situation we can expect to see is a situation in which mother fails to divide the strawberries. The impossibility of mother succeeding in the task explains why she fails every time. This impossibility is a consequence of only

the formal relationship between the number of strawberries and the number of children. No further details of the situation need to be considered to generate this result. The only details necessary to derive the explanandum are formal features of the model.[1]

Some philosophers of science have noticed explanatory patterns such as these and given them the name "distinctively mathematical explanations"[2] Typically, distinctively mathematical explanations are characterized as explanations that derive their explanatory force from appeal to formal mathematical facts. In the strawberry case, the presence of a mathematical fact (that 3 does not evenly divide 23) in the explanans would seem to qualify the explanation as distinctively mathematical. This is not entirely correct though. Although I do agree that the strawberry explanation does qualify as distinctively mathematical, we should not follow Mancuso (2008) in characterizing a distinctively mathematical explanation as one that is "carried out by essential appeal to mathematical facts" (p. 135). Just because an explanation appeals to a mathematical fact in the explanans does not mean the explanation is distinctively mathematical. If I had two children, this could be explained by the fact that I had one child, the fact that I had another, and the fact that 1+1=2. This explanation is not distinctively mathematical even though it appeals to a mathematical fact in the explanans. In this case, the facts doing the explaining are facts about me having children, not mathematical facts. Suppose though we narrow the explanandum and ask, given that I had one child and then another, why do I have two children instead of three? The explanation in this case is that, given I had one child and then another, I could not possibly have any other number of children than two.

1 Bokulich (2008) has identified what she calls "structural model explanations" (p. 40) which operate similarly to what I call constraining explanations.

2 c.f. Baker (2005), Batterman (2009), Lange (2013), Lyon & Colyvan (2003), Mancouso (2008) Pincock (2007, 2011, & 2015)

It is impossible that I have any other number of children because 1+1=2. In this case, the mathematical fact figures into the explanans by making it impossible that any other outcome could have occurred.

In order to differentiate between distinctively mathematical explanations and ordinary explanations that appeal to some mathematical fact, we must consider the precise way that the mathematical fact figures into the explanans. In the case with mother and her strawberries, the mathematical fact in the explanans constrains the possible explanandum that there can be. That 3 does not evenly divide 23 makes it so all possible scenarios are ones in which mother fails. In many of the examples of distinctively mathematical explanations that have been given in the literature, the mathematical fact in the explanans figures into the explanation by *constraining* the range of possible explanandum phenomena that one can expect to find. With this understanding of the role that mathematics can play in ruling out certain outcomes, all the distinctively mathematical explanations given in the literature can be seen as instances of constraining explanations, where unambiguously mathematical facts play the constraining role.

Consider this example, given by Lange (2013): Why does a double pendulum system have 4 equilibrium positions? A double pendulum system is just a pendulum hanging off of another pendulum, and an equilibrium position is a position that once placed in, the pendulum will remain there as long as the system is undisturbed. One way to answer this question is to consider the individual forces at play on the pendulum bobs, and then calculate the points at which both bobs feel a net force of 0. According to Newton's second law, if a system is placed in a configuration in which there is 0 net force

acting on that system, then the system will remain at rest as long as it is undisturbed. By calculating the forces working on the system, we can determine that the system has exactly 4 equilibrium points. This is causal explanation, as it involves identifying the specific forces acting on the pendulum bobs, identifying how they interact, and calculating at which position those particular forces cancel each other out (p. 502).

Lange points out that there is a "top down" explanation for this phenomenon too. A double pendulum system has a configuration space that can be represented as points on a toroidal surface. In other words, if we traced out every possible trajectory that a double pendulum system could take, we would get a donut-like shape. A torus is a surface with genus $g = 1$ (i.e. a sphere with $g = 1$ holes in it), and for any such surface (smooth, orientable or compactable, etc.) the number of minimum, maximum, and saddle points (which for our purposes represent equilibrium points) obeys the equation $N(min) - N(sad) + N(max) = 2 - 2g$, which is 0 for $g = 1$. Because the toroidal surface is compact, which means it is a closed set of points that are a fixed distance from each other, there is at least 1 maximum and 1 minimum, and thus, by the previous equation, 2 saddle points. Therefore, there are at least 4 equilibrium points in total.

Notice that this explanation does not work by describing any causal features of double pendulums. It does not work by describing the individual causal forces or mechanisms acting on the pendulum bobs, and it does not work by describing the specific physical constitution of any double pendulum system. All this explanation relies on is the fact that a double pendulum's configuration space is a torus, and Newton's second law of motion. This explanation applies to every possible double pendulum system too,

regardless the mass of the bobs and lengths of the strings. No matter what, any possible double pendulum will have at least 4 equilibrium points. There is no possible fact about the specific causes operating in any specific double pendulum system that could make it not true that the system has at least 4 equilibrium points.

This is another case of a constraining explanation. In the case of a double pendulum, the salient features of the model are the toroidal configuration space of a double pendulum system and Newton's second law of motion. In conjunction, these two features of the model necessarily determine that every possible double pendulum system will have at least 4 equilibrium points in total. In other words, the fact that the configuration space of a double pendulum is a torus constrains the possible manifestations of double pendulum systems to precisely those systems that have at least 4 equilibrium points. This explanation is decidedly non-causal, as it does not proceed by identifying causal entities, causal mechanisms, processes, causal histories, etc… The explanation identifies formal constraints on the tokens that are compatible with the model. Any token double pendulum is such that it will have at least 4 equilibrium points.

Notice that this explanation refers to a natural law: the law that any system that has 0 net force acting on it is at equilibrium. This is a specific case of Newton's second law of motion. Why does the inclusion of a law of nature not make this explanation an ordinary causal explanation? The answer is because Newton's second law does not describe any particular causal-force law. Rather, Newton's second law sets constraints on the mathematical form that any force law can take. Newton's second law, which is sometimes written as $\boldsymbol{F} = m\boldsymbol{a}$ (where $\boldsymbol{F}$ and $\boldsymbol{a}$ represent vectors with the same direction),

describes the basic constraints on how forces have to operate. Notice that other force laws in physics, such as Coulomb's law which represents the electrostatic force between two charged particles, also take this general form. Newton's second law, in a way, "transcends" (Wigner 1964, p. 995) the peculiarities of the individual forces there actually happen to be.

Indeed, the possibility that the individual force laws could have been different is an idea that has been entertained by physicists. Paul Ehrenfest in 1918 published a famous paper that demonstrated that if gravity were an inverse-cube force, or if the strength of gravity fell off at a greater rate over distance, the orbits of the planets would not be stable. Ehrenfest's argument requires that Newton's second law would still be true, even if the actual forces that populated our world were different. To put the idea in language more familiar to philosophers, in any possible world where there are forces, they must operate in such a manner that the force acting on an object is proportional to the mass and acceleration of that object. If any force law were formulated that did not conform to this constraint, we would reject it, not only because it would be empirically false, but because it would violate constraints on the valid forms that a force law can take.

Newton's second law has the interesting feature of coordinating and framing a group of individual laws; specifically force laws (as forces are formulated in classical physics). There are a certain subset of laws in physics that have this unique feature. Wigner calls these laws "invariance principles" and argues that they act as "touchstones for the validity of possible laws of nature" (1964, p. 997). Wigner further writes on the character of invariance principles as "rigorous correlations between those correlations

between events that are postulated by the laws of nature" (p. 997). The general idea is that invariance principles are kinds of "super principles" (p. 996) that coordinate and fix a framework in which natural laws exist. Richard Feynman also commented on a certain class of scientific principles which have this feature, writing "...across the variety of these detailed laws, there seem to be great general principles, which all the laws seem to follow" (1967, p. 59). Similarly, Lange (2009) & (2011) writes that symmetry principles in physics, which each represent particular invariant features of space-time, are a kind of "meta-law" that work by imposing constraints on the valid form of first order dynamical laws. Symmetry principles in physics are taken to be "laws that govern the laws governing subnomic facts" (2009, p. 110).

Physicists regularly appeal to invariant and unchanging features of a system in order to explain more specific features of that system. A classic example of an explanation using invariants is an explanation of the conservation of momentum in an isolated system. The Hamiltonian operator, which is a mathematical operator that expresses the total energy of a system, is invariant with respect to uniform translations. In other words, I can take a dynamical system, put it into uniform rectilinear motion, and the total energy of the system stays the same. It can be show mathematically that the laws of momentum conservation follow directly from this basic constraint on the form of the Hamiltonian operator. More interestingly, one does not need to make reference to any specific forces or causes acting in that system, and no specific laws of motion need to be consulted to get this result. The formal characterization of the Hamiltonian alone is enough to guarantee that the laws of momentum conservation hold. It is *because* the

Hamiltonian operator is an invariant quantity that the laws of momentum conservation hold, and this fact holds regardless of the specific content of the laws of motion, or the specific forces and causes working within the system. In this case, the invariance of the Hamiltonian operator explains the presence of conservation laws, because this invariant feature puts mathematical constraints on how dynamical systems have to behave.

This is an example of a constraining explanation, as the explanation does not work by identifying causes, mechanisms, or processes. Instead, the explanation works by identifying formal features of a particular model, in this case mathematical features of the Hamiltonian operator as it appears in classical mechanics, and shows hows the explanandum, the laws of momentum conservation, follow as a necessary consequence of the presence of these basic formal features. Any possible extra details that fill in the model must be consistent with those basic constraints.

Explanations involving invariance principles also play a significant role in the methodology of physics. Invariance principles, such as symmetries, are used to guide scientists in their search for natural laws.[3] For example, translational symmetry requires that any candidate natural law has such a form that it remains invariant under arbitrary distance translations: no matter where I move in space, the laws of nature should remain the same. Translational symmetry does not tell us about the specific content of natural laws, such as the exact strength of the gravitational force, but it does tell us about a general feature that all natural laws must have. Translational symmetry, which is a kind of invariance principle, constrains the set of possible natural laws to only the set of natural

3 This sentiment is expressed by Einstein's (1919) famous distinction between "constructive" and "principle" theories. Roughly, constructive theories seek to explain a phenomena by reducing it into its constiutive elements (ex. the kinetic-molecular theory of heat) while principle theories set constraints that other theories must follow (ex. special relativity's upper velocity of the speed of light *c*)

laws that have a form that is invariant under translational symmetry. This constraint, in turn, explains why the actual laws of nature are invariant under arbitrary uniform translations. This argument pattern generalizes as well, as demonstrated by Emmy Noether in 1915, who showed that for every differentiable symmetry in physics, there exists a corresponding conservation law (Kosmann-Schwarzbach 2011, p. 59).

2.2. Optimality Models in Evolutionary Biology

Constraining explanations are not limited to the highly formal and abstract models of modern physics. They also find a home in the special sciences. Chirimuuta (2017) considers a few examples of what she calls “efficient coding explanations” in computational neuroscience (p. 3). Roughly, efficient coding explanations aim to explain why certain groups of neurons have the information processing capacities that they do. These explanations do so by appealing to the efficiency of certain computational methods. Chirimuuta begins by noticing that analog computation and digital computation have particular trade-offs. Specifically, analog computation requires less energy, but is more prone to noise and error when precision needs to be high. Digital computation on the other hand is much more precise, but has a higher baseline energy cost. Next, Chirimuuta points out that the general optimal solution to maximize energy efficiency and minimize susceptibility to noise in a computational system is for that computational system to implement a hybrid form of computation (p. 11). A hybrid form of computation involves alternating between bouts of analog and digital processing. Small chunks of energy efficient analog computation can be interspersed with bouts of digital processing to ‘clean up’ the signal. The idea is that a hybrid computational system would take the

energy efficiency of an analog computing system and combine it with the precision of a digital computing system.

The trade-offs between energy efficiency and noise susceptibility that are found in analog and digital computing systems do not seem to be a result of specific physical facts about computing systems. By definition, the individual components of a digital system can only represent one bit of information at a time (ex. 1 or 0 in binary). In contrast, the information content of an analog signal varies continuously over time with respect to some continuous variable (ex. voltage). For any given signal, a digital system needs more components to represent that signal than an analog system does. A digital system needs at least two components to represent 2 bits of information, but the same amount of information could be expressed in an analog system with a single component, provided that four different signals could be associated with four different states of that component. It follows that analog systems are less hungry for resources, but the more information one tries to encode in an analog system, the more susceptible to error the system will be. The more information content one tries to encode in a signal in a single component of an analog system, the range of physical states of that component that can be unambiguously associated with that signal becomes smaller. These facts about digital and analog computation are true regardless of the actual physical nature of the components of the computational system, be they metal wires, axons, sticks, or rocks. The respective trade-offs that analog and digital computation have are a result of the mathematical definitions of analog and digital computation, not empirical facts about analog or digital computing systems.

This account of hybrid computation is then applied to the neurological structure of the brain. Flow of electrical signals through the dendrites resulting in the firing of action potentials in axons is, Chirimuuta argues, remarkably like the structure of a hybrid computing system. Information in the dendrites is processed in a linear manner, much like analog computation, and action potentials are discrete, all-or-nothing events, much like digital computation. Other authors such as Dogulas et. al. (1994), Sarpeshkar (1998), and Clark et. al. (2006) have also suggested that information flow in the axons mimics a hybrid computational system. Chirimuuta argues that the neurological structure of the brain is explained by the efficiency of the hybrid computational system it implements. Specifically, hybrid computation is the optimal general solution to maximizing energy efficiency while minimizing noise in a computational system. In the context of natural selection, we can expect evolution to select for resource management strategies that are more efficient. Thus, the efficiency of hybrid computation explains why the particular neurological structure of the brain was selected for, and thus, explains the actual neurological structure of the brain.

This particular explanation makes use of evolutionary assumptions about selection and optimization of traits via the process of natural selection. Evolutionary biologists frequently use what are called *optimality models* to explain why organisms have the particular adaptations they do (Rice 2015, p. 589). Optimality models make use of mathematical optimization theory to answer questions about which biological strategies are most conducive to survival and reproduction in an environment. Optimality models identify *design* variables that needs to be optimized, and describe the *control* variable that

will optimize the design values. The model generates a *strategy set* and connects each possible strategy to some value of the design variable. Built into these models is information about the inherent constraints and trade-offs between the control variables (p. 589). For example, if we are building a bridge across a river, there are many design features that need to be optimized (length, width, height, etc.). But we cannot optimize theses features all at the same time. Certain trade-offs (more height = more money) and limitations (amount of money) constrain what the optimal design for our bridge can be. In some cases, we can overcome constraints and trade-offs between variables; for example we could accrue more funding for our bridge project and thus make a bigger bridge. Some constraints are more difficult to overcome. For example, it is harder to over come the constraint that the width of the river puts on the possible designs of the bridge. Once we have specified the control variables, how those variables optimize the design variable, and what the relevant constraints and trade-offs between the control variables are, we can deduce what strategy is best for optimizing the design value, and why that strategy is the best available solution.

The scarcity of resources in the evolutionary world puts a hard constraint on the possible strategies for survival that evolved computational systems can utilize. Evolutionary systems that implement computations need to use as little energy as possible, yet still be capable of accurately processing amounts of complex information so they can react appropriately to their environment. As such, one design value that is important for evolved computational systems is that they maximize energy efficiency and minimize error from noise. The relative trade-offs between efficiency and noise in analog

and digital computing systems will be represented as control variables in the optimality model. We could then use the model to determine that hybrid computation, all other things being equal, maximizes the ratio of energy efficiency to noise, given the environmental limitations. Thus, the model explains why the human brain has a particular neurological structure; because that neurological structure is the best available strategy that solves a particular optimization problem.

Another example given by Lyon & Colyvan (2003), involves bees and their hexagonal honeycomb tiling. Lyon & Colyvan ask why is it that bees always tile their honeycombs in a hexagonal grid. The explanation is that hexagonal tiling is the optimal way of tiling a 2-D plane into equal partitions, while minimizing the perimeter, a conjecture that was proven by Thomas Hales in 1999[4]. This fact about hexagons in conjunction with the fact that bees that make such a tiling pattern have more energy to contribute to survival and reproduction, explains why bees make hexagonal honeycombs. This explanation also works by noting that hexagonal honeycombs solve a particular optimization problem. The hexagonal shape of the honeycombs maximizes the ratio of the quantity of an area covered to energy needed to cover that area. In much the same way, Chirimuuta's explanation of the neurological structure of the brain makes reference to a more or less mathematical fact about information theory, and an evolutionary fact about selection pressures. Hybrid computation solves a resource management problem, and in the long run, natural selection can be expected to select for more optimal resource management strategies.

4 As such, Hale's conjecture is sometimes called the "honeycomb theorem"

Optimality explanations, such as the one given for the neurological structure of the brain, show that the initial conditions and the actual selection forces at play could have been different, but the explanandum phenomenon still would have occurred. Given that some facts are held constant, such as the scarcity of resources in the evolutionary world, it is likely that the explanadum phenomenon would have still occurred given a range of possible different causal antecedents. A causal explanation may show us how the phenomena is produced from prior initial conditions, but an optimality explanations shows us how the phenomena would still have occurred, even given a number of different causal antecedents.[5]

I argue that optimality explanations, such as the one given about the hybrid computational structure of the brain and the hexagonal shape of honeycombs, are a kind of constraining explanation. The trade-offs between the control variables in optimality models often cannot be understood as causal relations obtaining between mechanisms in a specific biological system. Instead, the trade-offs in optimality models can be understood as representing generic formal dependencies holding between functional features of a whole class of systems, quite independent of the actual causal details that characterize the systems. The relative trade-offs of analog and digital computation are not the result of the physical constitution or causal interactions of the components of a particular computational system. Analog and digital computation systems can be realized in mechanistically distinct ways, yet in all of these realizations, we would still see the relative trade-offs obtaining. When we plug these trade-off into an optimality model, the

5 Batterman (2002) has presented a class of models in physics he has characterized as minimal models, which also share this indifference to particular causal antecedents.

possible strategies that optimize the design value are constrained to those possible strategies that involve hybrid computation. Moreover, this explanation tells us why we should expect that the brain would have implemented hybrid computation, given a number of different possible causal histories of the evolution of the brain. We expect there to be strategies that involve hybrid computation because the model constrains the viable strategies to those strategies that optimize a certain design value, and hybrid computation optimizes that value. The specific causal history of the selection process makes relatively little difference to the occurrence of the explanandum, thus the explanation cannot work by primarily citing causes, mechanisms, or processes. Instead the explanation works by constraining the possible causes, mechanisms or processes we can expect to see at work in the brain. While we are not constrained to any one picture of the brain, we are constrained to the disjunction of those possible strategies that involve the brain implementing some form of hybrid computation.[6]

Recall the previous discussion about constraining explanations in physics. Constraining explanations in physics work by finding some quantity or feature in a model that is invariant with respect to some changes, and deriving from that invariant feature other general properties that the model must also have. In a sense, this explanation for the efficiency of the human brain makes reference to invariant features about information transfer. Sarpeshkar (1998) describes the relative pros and cons of analog and digital computation by defining “resource precision curves” (1998, Figure 3). These mathematical graphs show the relationship between resource consumption and signal to

6 Rice (2015) makes a similar point about the disjunctive nature of “equilibrium explanations” in biology (p. 596).

noise ratio for both analog and digital computation. What they show is that, while the costs of analog computation remain low when the precision is low, once precision reaches a certain limit, energy consumption for analog computation increases rapidly by orders of magnitude. In contrast, digital computation has an initially high rate of energy consumption, but the energy costs remain comparatively low, even when precision is high.

The particular equations represented in the graphs presented by Sarpeshkar were generated by making concrete assumptions about parameters, such as the length of the transistors, the kind of transistors, and the noise distribution of the signal in a range of physical substances. The question is, as Chirimuuta rightfully points out, whether these resource precision curves would take the same form for any particular value of the parameters (2017, p. 16). If so, then the resource precision curves for analog and digital computation could plausibly constitute a kind of invariant property of analog and digital computation. More specifically, the invariant property is the ratio of efficiency to signal/noise ratio of analog and digital computation. These invariant features of analog and digital computation are essential to understanding why hybrid computation, all other things being equal, maximizes the efficiency to noise ratio of a computational system. This would be true for any signaling system.

Much like spatio-temporal symmetries and conservation laws work by constraining the possible forms that natural laws in physics can take, optimality models in biology work by constraining the forms that evolutionary phenomena can take. For any possible information processing systems, the viable strategies that optimize the efficiency

to noise ratio of that system will be constrained to those outcomes in which the system implements a form of hybrid computation. Most importantly, this fact is not explained by referring to the specific physical or causal features of any particular computational system, or even the specific causal selection pressures that are at play.

The main difference between constraining explanations used in physics and those used in the biological sciences is the degree of stability and invariance that the explanation gives the explanandum. In the case of explanations involving invariance principles, the explanandum phenomena, such as conservation laws, are a very stable phenomena. Conservation laws would still exist given a significantly large range of different possible ways the world could be. In fact, it seems that the only way the conservation laws would be different is if the fundamental nature of space-time were different. In this sense, conservation laws are an extremely stable phenomena. In contrast, evolutionary phenomena is not nearly as stable. Obviously, evolutionary phenomena could have easily not occurred given a different causal history of the world. A meteor could easily have wiped out all life on earth, or the composition of earth's atmosphere may have been different and thus not allow the development of complex, oxygen and carbon based life. However, *given that* the earth developed in such a way that evolutionary phenomena could exist, optimality explanations give evolutionary phenomena a large degree of stability and invariance. Given that evolutionary phenomena exists, biological structures that optimize design constraints are very likely to be found in nature.

That claims of stability and invariance are relative to background considerations reflects a simple fact about explanation. In general, our explanatory goals are relativized to different ends. Whether something counts as explanatorily relevant depends upon our goals and interests. Imagine I ask someone "Why does evolutionary phenomena P exist?" and they answer "Because the big bang happened." In one trivial sense they are correct. Had the big bang not occurred, evolutionary phenomena P would not exist. Of course, in most contexts, we would not count this as a legitimate explanation of evolutionary phenomena P. Facts about the big bang are irrelevant to facts about evolutionary phenomena.[7] Biologists do not go around taking cosmological facts about the big bang into account when formulating explanations for biological phenomena. In the domain of biology, some facts are taken for granted, such as the existence of evolution, and this allows biologists to focus on explaining more specific phenomena, such as the existence of specific biological traits.

3. Beyond Causal Understanding

One unique feature of constraining explanations is that they give us a kind of modal knowledge that causal explanations cannot furnish. Constraining explanations show us how the occurrence of a phenomenon is, in a sense, inevitable or necessary. This inevitability or necessity is relativized to the domain of interest. Constraining explanations show how a given explanandum would have occurred, even given a number of different possible causal antecedents. Constraining explanations can give us this knowledge because they highlight general necessary features that all possible models of

7 I must limit this claim slightly. It is conceivable that in some cases, facts about the big bang may be directly relevant to the occurrence of evolutionary phenomena, such as facts about the mass-energy distribution of the universe following periods of inflation. However, in general, considerations about the big bang are not treated as relevant to explaining evolutionary phenomena.

the target system must have. The necessary general features are determined by the mathematical formalization of the system of interest.

The last paragraph is hard to understand in the abstract, so here is an example of a real world phenomenon that was once explained in causal-mechanical terms and is now explained in constraining terms. Let us consider the Lorentz transformations, the mathematical equations that govern coordinate transformations between different inertial frames of reference. The explanandum of interest is the obtaining of the Lorentz transformations. Why are inertial frames of reference related to each other by means of these mathematical transformations? Lorentz himself attempted to explain the presence of the transformations in causal-mechanical terms. Precisely, Lorentz was guided by the assumption that matter is held together by electromagnetic forces, forces that are propagated in the aether. So, when objects are moving very quickly relative to the aether, the forces become distorted, and the object contracts in the direction of motion.[8] Lorentz believed this contracting process is the causal phenomena that the mathematical transformations are supposed to represent. This is a causal explanation of the Lorentz transformations, as it appeals to the specific mechanical and electromagnetic forces operating on the object, and to how those causal elements interact to produce the contracting behavior.

In contrast, let us look at the explanation of the Lorentz transformations as given by modern physics. Modern physics explains the Lorentz transformations in terms of spatio-temporal symmetries, the principle of relativity, and the invariance of the space-

8 Incidentally, this is also why the Lorentz transformations are sometimes called the Lorentz ‘contractions’

time interval between two events. Avoiding the technical mathematical derivations, we can derive the presence of the Lorentz contractions simply from constraints on the mathematical model of the theory.[9] This is clearly NOT a causal explanation. As Lee & Kalotas (1975) are quick to point out, the presence of the constant *c* in the mathematical derivations of the transformations should be understood as an arbitrary constant with the dimensions of velocity (p. 436). In particular, we should not interpret *c* as referring to a specific speed of light, or any particular physical process. To do so would be to suggest that the existence of the Lorentz transformations depends on a particular causal process, which is not the case. Spatio-temporal symmetries and the principle of relativity do not describe causal relations, mechanisms, or processes. Rather, spatio-temporal symmetries and the principle of relativity represent invariant features of space-time; they put formal constraints on what kinds of events or interactions can occupy that space-time.

This explanation gives the explanandum a particular sort of inevitability or necessity that is stronger than that which ordinary causal explanation could bestow upon. In the case of the causal-mechanical explanation of the Lorentz contractions, the explanandum is a consequence of the specific causal forces acting on the object. The existence of the contractions is dependent on specific causes, i.e. the electromagnetic forces acting on the object. If these causes were different; let's say the strength of the electromagnetic attraction were stronger, then the Lorentz transformations would have a different form. In contrast, according to modern physics, the existence of the Lorentz transformations follow as a necessary consequence of constraints on physical measurement. The Lorentz transformations would be present, regardless of the specific

9 For a more technical derivation, refer to Lee & Kalotas (1975).

laws or causal interactions that obtain in that framework. The existence of the Lorentz transformations is neither sensitive to, nor dependent on the actual causes, mechanisms, or processes that occupy space-time. The existence of the Lorentz transformations is prior to the existence of any causes, and would still be present even if the actual causes that populated space-time were very different.[10] If we are to understand explanation as Bokulich (2008)[11] does as consisting of identifying the counterfactual dependencies that hold in a model, then the Lorentz transformations cannot have a causal explanation. Counterfactual differences the describe scenarios in which the causes that actually populate the world are different makes no difference to the existence of the Lorentz transformations. They would still be there, even if the force of gravity were a bit stronger, or if the strong nuclear force operated over a large distance, or even if there were only a single lonely electron in the universe.

The key feature of constraining explanations is that they show us that things are a particular way because they *could not be* any other way. It is *impossible* that we will find a double pendulum system that has less than 4 equilibrium points. This impossibility explains why we have never actually found, or successfully created, a double pendulum system that has less than 4 equilibrium points. Likewise, given formal constraints of special relativity, it cannot be that the Lorentz transformations do not occur. Many accounts connect correct explanation with a kind of expectability (Salmon 1984, Batterman 2002, Strevens 2008, Rice 2015). That is, an explanation explains partly in

10 Lange (2016, p. 155) makes a similar point about the universality of the parallelogram force law for vector addition.

11 Bokulich, like other philosophers such as Pexton (2015), finds it useful to decouple the causal interventionist elements from the counterfactual explanatory elements of Woodward's (2003) manipuability theory of causation.

virtue of showing how the explanandum was to be expected. The key feature of constraining explanations is that they show us a particular explanandum was to be expected, regardless the specific causal antecedents that precede the explanandum. We can use this knowledge to make predictions as well. Formal principles in physics allow me to predict that any newly discovered law of nature will have a particular mathematical form. Reasoning from constraints in the biological realm allows me to predict general functional features that evolved systems are likely to have.

If a model is understood like a game, constraining explanations allow us to find those events, objects, states of affairs, etc. that do not abide by the rules of the game and toss them out from the get go. Say we are playing chess. Just from looking at the rules of chess, we cannot determine what the actual positions of the pieces are. However, just from looking at the rules of chess we can determine which positions the pieces *cannot* be in. If I am playing white, then I know that my right hand bishop will never be on a black square (assuming I am competent at chess and not cheating). The rules of chess preclude that my right hand bishop could ever be on a black square. Consequently, this restriction explains why it is that my right hand bishop is always on a white square. My right hand bishop is on a white square because the rules of chess constrain the possible positions that it can be in. Throwing out all the possible phenomena that do not abide by the rules explains why all the possible phenomenon left (aka, the ones we can expect to actually come across) do abide by the rules of the game. If all possible outcomes that are consistent with a particular model have a particular feature, then that explains why the actual outcome has that specific feature.

Constraining explanations can work side by side with other more causal approaches to explanation. The relevant formal features of the model end up constraining the possible kind of causes or mechanisms we could posit to explain phenomena.[12] The fact that the speed of light is invariant in all reference frames give me some information about the causal mechanisms I can expect to find in the world. At the very least, it tells me that no possible causal mechanism will involve non-local influences faster than the speed of light. Also, it tells me that I should not go around positing superluminal mechanisms to explain things. In the case of optimality explanations in biology, the mechanisms that realize the system that optimizes the desired parameter value have to have the right kinds of properties. At the very least, the possible causal mechanisms have to be structured in such a way that they optimize the required parameters.

Some proponents of causal theories of explanation do not agree that merely constraining the space of possible mechanisms has explanatory significance. For example, Craver (2008) argues that Hodgkin and Huxley's mathematical model of the neuronal action potential is not explanatory, because the equations for the model "failed to appreciably constrain the space of possible mechanisms for the conductance changes" (p. 1026). According to Craver, Hodgkin & Huxley's mathematical model fails to explain because it gives one "no reason to privilege this one how-possibly model above the others as a how-plausibly or how-actually model" (p. 1029). Craver's argument seems to be that the model is not explanatory on its own, because the equations in Hodgkin & Huxley's model reveal only extremely general properties of the possible mechanisms that may be

12 One could see constraining explanations as narrowing the set of "how-possibly" (Craver 2006, p. 6) models that we can consider; that is, sketches of the possible mechanisms that are responsible for a given phenomena.

involved. This generality is considered a weakness of the model. A complete explanation would involve a detailed model of the actual physical mechanisms in the brain that constitute the action potentials.

It should be mentioned that when dealing with extremely complex physical systems such as the brain, narrowing down the space of possible mechanisms that may be in play is no trivial matter. The brain is the most complex machine we know, so even identifying extremely general features and properties of the mechanisms at play in the brain is a significant achievement. Hodgkin & Huxley's model determined that the mechanisms in play in the action potential have to be such that they can realize electric currents and transmit electric charges in the way that the model specifies. This is a significant explanatory insight into the working of the action potential. Craver's main claim seems to be that the model does not explain because it does not tell us how the action potential *actually* works. But the model does tell us this; action potentials *actually* work by storing and transmitting electric charges and currents. The electric potentials and currents that the model represents are extremely explanatorily relevant to the working of the action potential.[13] The explanatory importance of the dynamical equations of the HH model of the action potential is one of the reasons that the search space for possible mechanisms of the action potential is constrained only to those mechanisms that can realize the dynamical equations. The constraint that the dynamical model places on the search space for possible mechanisms in the brain plays an invaluable role in theorizing and constructing possible mechanisms of the action potential.

13 This point is also argued for by Levy (2013).

Moreover, the insistence that models are most explanatory when they describe causal relations between elements of actual mechanisms, as is the case with Kaplan (2011)'s 3M account, may not be the whole story. Models in science are supposed to have a certain general character. They explain by representing the generic properties of a class of systems that are explanatorily relevant to the behavior of that class of systems. These generic properties can often be represented independently of the actual mechanistic details of a system. In some cases, models explain even when intentionally misrepresenting causes or mechanisms. Batterman (2009) points out various explanations in the physical sciences that deliberately introduce false representations and non-actual idealizations to explain. The universality of critical phenomena in various systems, such as phase shifts in chemical substances, is explained by relating models via limit transformations performed on parameters of structural equations. These mathematical limit operations deliberately introduce limiting idealizations into models to explain the nonanalytic, qualitative changes in the macroscopic behavior of a system that occur at the critical points of that system (Batterman 2009, p. 7). Mathematical idealizations do not represent causes or mechanisms, yet scientists routinely invoke them to explain, which is in direct tension with the 3M mapping account of explanatory models.

One universally agreed upon virtue of a good explanation is its scope and applicability. In describing explanations as such, I am explicitly drawing upon unificationist accounts of explanation, such as that given by Kitcher (1989). A good explanation should be applicable to a number of different circumstances. Newtonian mechanics was hailed as so incredibly explanatory precisely because it unified terrestrial

and celestial mechanics under a single overarching framework. Likewise, fields in the special sciences, such as molecular genetics, or computational neuroscience, are hailed as explanatory because they succeed in uniting distinct classes of phenomena into a single explanatory framework. This generality is not a weakness of explanation. We can see how the behavior of systems that initially seem to have relatively little in common can be explained by using the same general principles.

4. Conclusions

I have argued that constraining explanations are a distinct kind of explanation in science, different from causal explanation. Constraining explanations work by identifying formal features of models and demonstrating that the explanandum was to be expected solely on the basis of those formal features. These explanations do not proceed by identifying causes, mechanisms, or processes. Instead these explanations work by identifying formal constraints on the possible events, objects, states of affairs, causes etc that there can be. Constraining explanations are unique because they tell us that a particular phenomenon, in a sense, just had to be that way.

It should be stressed that I am not arguing that causal approaches to explanation are not a legitimate approach to explanation. In a large number of circumstances, explanations involving causes, mechanisms, or processes are the right kind of explanation. I am arguing though that causal-mechanical approaches are not the *only* legitimate approach to explanation. Causal-mechanical models are just one kind of explanatory style. In my view, it is unlikely that there is a one-size-fits-all approach to scientific explanation. A commitment to a monist account of explanation runs the risk of

missing the diverse range of explanatory patterns that are actually present in the sciences. A commitment to a pluralism of explanatory styles in the sciences opens up new avenues of philosophical research. Explanation is a multi-faceted entity and each different kind of explanation contributes to our knowledge of the structure of the world.

Works Cited

Baker, A. “Are There Genuine Mathematical Explanations of Physical Phenomena?” *Mind.* Vol. 144. No. 454. (Apr. 2005). pp. 223-238.

Batterman, R. “Asymptotics and the Role of Minimal Models” *The British Journal for the Philosophy of Science.* Vol. 53. No. 1. (Mar, 2002). pp. 21-38.

Batterman, R. “On the Explanatory Role of Mathematics in Empirical Science” *The British Journal for the Philosophy of Science.* Vol 61. No. 1. (March 2009). pp. 1-25.f

Bokulich, A. “How scientific model can explain” *Synthese.* Vol. 180. (2011). pp. 33-45.

Chirimuuta, M. “Explanation in Computational Nueroscience: Causal and Non-causal” *The British Journal for the Philosophy of Science.* No. 0 (2017). pp. 1-32.

Clark, B.; Häusser, M. “Neural Coding: Hybrid Analog and Digital Signalling in Axons.” *Current Biology.* Vol. 16. No. 15. (August, 2006). pp. 585-588.

Craver, C. “When mechanistic models explain.” *Synthese.* Vol. 153. No. 3 (December, 2006). pp. 355-376.

Craver, C. “Physical Law and Mechanistic Explanation in the Hodgkin and Huxley Model of the Action Potential” *Philosophy of Science.* Vol. 75. No. 5 (December, 2008), pp. 1022-1033.

Craver, C. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press. (August, 2009).

Douglas, R.; Mahowald, M.; et. al. “Hybrid Analog-Digital Architectures For Neuromorphic Systems.” *Neural Networks. IEEE World Congress on Computational intelligence.* Vol. 3. (1994). pp. 1848-1853.

Einstein, A. “What is the Theory of Relativity?” *The London Times.* Nov. 28, (1919)

Elster, J. *Explaining Technical Change.* Cambridge University Press. (1983)

Hempel, C; Oppenheim, P. “Studies in the Logic of Explanation” *Philosophy of Science.* Vol. 15. No. 2. (Apr., 1948). pp. 135-175.

Kaplan, D. “Explanation and description in computational neuroscience.” *Synthese.* Vol. 183. *Synthese.* (2011). pp. 339-373.

Kitcher, P. “Explanatory unification and the causal structure of the world.” in *Scientific Explanation* (eds. Kitcher, P. Salmon, W.). University of Minnesota Press. (1989). pp. 410-505.

Kosmann-Schwarzbach, Y. *The Noether Theorems: Invariance and Conservation Laws in the Twentieth Century.* Springer. (2011).

Lange, M. *Laws and Lawmakers: Science, Metaphysics, and the Laws of Nature.* Oxford University Press. (2009).

Lange, M. “Conservation Laws in Scientific Explanations: Constraints or Coincidences?” *Philosophy of Science.* Vol. 78 (July, 2011). pp. 333-352.

Lange, M. “What Makes a Scientific Explanation Distinctively Mathematical?” *The British Journal for the Philosophy of Science.* Vol. 64. (2013). pp. 485-511.

Lange, M. *Because Without Cause: Non-causal Explanation in Science and Mathematics.* Oxford University Press. (2016).

Lee, A; Kalotas, T. “Lorentz transformations from the first postulate.” *American Journal of Physics*. Vol. 43. No. 5. (1975). pp. 434-437.

Lyon, A.; Colyvan, M. “The Explanatory Power of Phase Space” *Philosophia Mathematica.* Vol. 16. No. 2. (June, 2008). pp. 227-243.

Mancosu, P. “Mathematical Explanation: Why it Matters” in *The Philosophy of Mathematical Practice.* Oxford University Press. (2008). pp. 134-150.

Pexton, M. “There are non-causal explanations of particular events” *Metaphilosophy.* Vol. 47. No. 2. (Apr., 2016). pp. 264-283.

Pincock, C. “A Role for Mathematics in the Physical Sciences.” *Nous.* Vol. 41. (2007). pp. 253-75.

Pincock, C. “Mathematical Explanations of the Rainbow” *Studies in the History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics.* Vol. 42. No. 1. (2011) pp. 13-22.

Pincock, C. “Abstract Explanations in Science” *The British Journal for the Philosophy of Science.* Vol. 66. (2015). pp. 857-882.

Rice, C. “Moving Beyond Causes: Optimality Models and Scientific Explanation.” *Nous.* Vol. 49. No. 3 (2015). pp. 589-615.

Salmon, W. *Scientific Explanation and the Causal Structure of the World.* Princeton University Press. (1984).

Salmon, W. “Causality without Counterfactuals.” *Philosophy of Science.* Vol. 6. No. 2. (Jun., 1994). pp. 297-312.

Sarpeshkar, R. “Analog versus digital: extrapolating from electronics to neurobiology.” *Neural Computation.* Vol. 10. No. 7. (October, 1998). pp. 1601-1638.

Strevens, M. *Depth: An Account of Scientific Explanation.* Harvard University Press. (2008).

Woodward, J. *Making Things Happen: A Theory of Causal Explanation.* Oxford University Press. (2003).

Wigner, E. “Events, Laws of Nature, and Invariance Principles” *Science.* Vol 145. No. 3636. (Sep., 1964). pp. 995-999.