Forthcoming in Minnesota studies in the philosophy of science

The Causal Economy Approach to Scientific Explanation

L.R. Franklin-Hall

Abstract

This paper sketches a causal account of scientific explanation designed to sustain the judgment that high-level, detail-sparse explanations—particularly those offered in biology—can be at least as explanatorily valuable as lower-level counterparts. The motivating idea is that complete explanations maximize *causal economy*: they cite those aspects of an event's causal run-up that offer the biggest-bang-for-your-buck, by *costing less* (in virtue of being abstract) and *delivering more* (in virtue making the event stable or robust).

keywords: causal explanation, black-boxing, abstraction, stability, robustness, explanatory level, explanatory trade-offs

1—Introduction

What distinguishes good scientific explanations, those that enlighten and captivate us, from their work-a-day peers, which may minimally account for the phenomena at hand but without inspiration? *Abstraction* is a popular answer. One explanation is more abstract than another when its explanans says less than does the other, ruling out fewer ways that the world might be. Since high-level explanations—those exploiting the special vocabularies of economics, psychology and biology, for example—are informationally impoverished compared to those offered, at least in principle, by basic physics, it is unsurprising that abstraction has been important to those who aspire to make sense of the special virtues of explanations provided by the high-level sciences.

But just why is abstraction such a good thing? Two distinct strategies have been deployed to account for abstraction's special worth: appeal to the explanatory value of generality and to the importance of causal difference-makers.

Pursuing the first path, Putnam (1975) claimed that explanations citing abstract geometrical properties were optimal in virtue of applying to a wide range of different actual cases, such as to similarly shaped pegs and holes made of different

materials. Kitcher (1981, 1984, 1999) too accounted for abstraction with the help of the more fundamental virtue of generality. He argued that correct explanations appeal to cohesive argument patterns that are actually instantiated in a wide range of systems, and that any pattern doing so will necessarily abstract from the "gory details" (Kitcher 1984: 370). And following Levins (1966), Sober (2000) suggested that abstraction results from the need to generalize, writing that "in order to isolate general patterns, we abstract away from the idiosyncrasies that distinguish some objects from others"(66).

Generality-based (or 'unificationist') explanatory approaches find fewer adherents today, and causal accounts are increasingly dominant. Though early causal theories, such as those offered by Railton (1981) and Salmon (1984), made no room for abstraction, currently popular difference-making accounts do better. Because they hold that non-difference-making causal influences are *not* explanatory, they are able to recommend at least some measure of abstraction without relying on the explanatory value of generality itself. For instance, on Strevens' (2008) kairetic view, to explain an event is to cite a causal model that has undergone an optimizing procedure that leaves it including a set of relatively abstract, but still physical, difference-makers. And Woodward (2003, 2010) suggests that his interventionist account can also support the judgment that high-level, abstract explanations are, at least sometimes, optimal.

Though all of these explanatory approaches are rich in insights, they each leave something to be desired as accounts of high-level explanation. Speaking against generality theories is the widespread hunch that generality in itself is explanatorily irrelevant, a mere byproduct of explanations that cite causes rather than an independent source of explanatory force. A second problem is that, in spite of the way they are advertised, if the most general explanatory patterns are provided by the physical facts and laws, generality-based theories may in fact leave no room for high-level explanations. Of course, this conclusion might be resisted by distinguishing the good or 'real' patterns, under which phenomena should be unified, from the explanatorily irrelevant patterns, which are pitched at levels either too high or too low or that that seem gerrymandered. But this distinction has proved difficult to draw in a principled way. For instance, in perhaps the most influential defense of high-level biological explanation to date, Kitcher (1984) simply appeals to a brute and unexplained notion of processes that constitute *natural kinds* to do the work.

Causal-difference-making accounts, though avoiding some of the difficulties just mentioned, also face problems, among the most notable of which is this: though they permit some level of explanatory abstraction, no extant causal theory recommends explanations nearly as abstract as those that high-level scientists actually formulate. For instance, on Woodward's (2003) basic interventionist view, an explanation is explanatory in virtue of answering what-if-things-had-been-different questions. High-level explanations can answer some such questions, and are thereby somewhat explanatory. But fully reductive accounts would seem to answer more, and so Woodward's theory appears committed to regarding them as superior.[1]

The account of scientific explanation outlined here—the *Causal Economy* account—aims to go much further than competing theories in recognizing explanations as abstract as those that high-level scientists actually provide as complete, optimal explanations. Moreover, it does so within a causal framework that eschews any direct appeal to generality or to metaphysically questionable notions like top-down causes, emergent properties, patterns, high-level individuals, or high-level natural kinds.

To set the stage for this view, I proceed in section 2 to present an example of *explanatory black boxing*, a pervasive style of explanation in the high-level sciences. Section 3 elaborates on the difficulties facing proponents of causal explanation who aim to make sense of abstract explanations of this kind. In section 4 the Causal Economy account is outlined and in section 5 its details are described. Section 6 considers how the account treats simple examples and might be adapted to more complex ones. Section 7 concludes by exploring the basic rationale for the entire Causal Economy explanatory picture.

2—Motivating Example

[1] This interpretation of the commitments of the interventionist approach is particularly clear in Hitchcock and Woodward (2003). Woodward (2010) has more recently suggested that a distinct explanatory standard, proportionality, can also support a high-level preference. My [2016a] argues that this particular proposal is not successful. For a recent attempt to avoid such reductionist implications by using the interventionist approach in combination with information theory, see Andersen (2017).

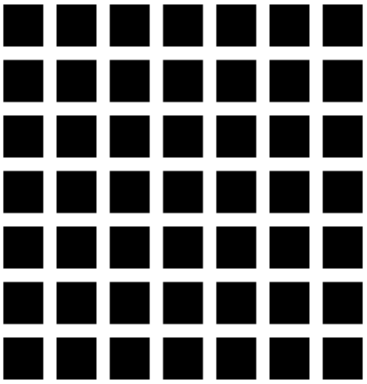
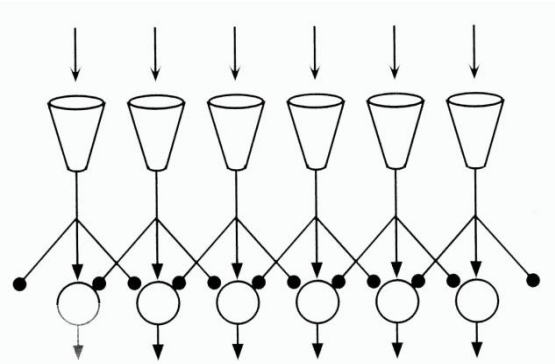


Figure 1: The Hermann Grid

Printed above is a Hermann Grid. While the background between the black squares is uniformly white, most people see ghostly gray smudges at the intersections of the squares. How might we explain the particular event of my perception of these smudges right now? The classic account of this illusion, offered first by Baumgartner (1960), goes as follows:[2] the receptive field of my eye, the retina, includes two layers of neurons: a layer of light-receptive elements (photoreceptors) and a layer of retinal ganglion cells, one abutting the other. They are connected in a lateral inhibitory network, where retinal ganglion cells are activated by the firing of photoreceptors at the same location in the layer as themselves, and inhibited by the firing of neighboring photoreceptors on all sides. This pattern of connections is illustrated in cross section in the diagram below (figure 2).



[2] Though pervasively cited in textbooks (Sekuler and Blake 1994, Palmer 1999, Snowden 2006), this account is not uncontroversial (Spillman 1971). Other explanations have been offered (de Lafuente and Ruiz 2004; Ash, Comerford, and Thorn 2003), though no particular alternative has gained wide support. I set aside these controversies here, since the classic explanation would be a good one, if true, and since the rival explanations (invoking cortical mechanisms) would also involve 'black-boxing,' which is the feature of philosophical interest here.

Figure 2: Cross-section of lateral inhibition network (modified from Palmer 1999: 116). Photoreceptor cells are represented by cones; retinal ganglion cells by circles. Excitatory relations are indicated by arrows; inhibitory relations by dots.

These lateral inhibitory connections between my neurons, in combination with the input pattern of light to my photoreceptor cells, provide the foundation for an explanation of the appearance of grey spots at the intersections of the Hermann Grid. When I look at the grid, light focused on my retina takes the pattern of the shape printed on the page. Photoreceptor firing frequency is in proportion to light exposure. And given the connective architecture already described, photoreceptor firing activates retinal ganglion cells immediately adjacent to them and inhibits their neighbors. Retinal ganglion cells corresponding to photoreceptors that receive light from some place along the non-intersecting white areas of the image are only laterally inhibited by neighboring excited cells on *two* sides, while cells at intersections are inhibited by neighbors on *four* sides (see figure 3). This heightened inhibition decreases the firing rate of retinal ganglion cells corresponding to the intersections at which gray smudges are seen. Since there is a direct relationship between ganglion firing rate and the brightness of the white perceived, we can in this way account for the perception of gray smudges at intersections in terms of light input and the particular network architecture.

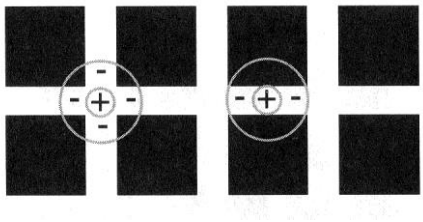


Figure 3: Illusion explained (from Palmer 1999: 118)

3—The Black-Boxing Challenge

Many scientific texts offer versions of the above explanation. And it certainly seems illuminating, arguably describing just the features relevant to the target phenomenon, without extraneous molecular details. Furthermore, in form it is typical of explanations found elsewhere in neuroscience, developmental and

evolutionary biology, and across the high-level sciences.[3] For instance, developmental biologists offer explanations structurally similar it (i.e., invoking lateral inhibitory networks) in accounting for how certain cells are determined to become hair cells in the fruit fly, or how certain cells are determined to become different neuronal types in zebra fish (Appel et al. 2001).

Yet it is exceedingly difficult to make sense of the special value of the above explanation from a strictly causal perspective. However the details might be spelled out, a central insight of the causal approach is that explanations *show how things work*. The Hermann Grid explanation does show *something* about how things work, for instance by describing the effects of photoreceptor firing on ganglion cells. But it says much less about how things work than is currently possible, or (some might think) than is in principle desirable.

In particular, it says less than it might in virtue of its *black boxing*—that is, in virtue of the fact that, rather than citing physical laws and physical arrangements to account for the target event, it appeals to a few input conditions in concert with a series of interconnected entities (the *black boxes*) understood to produce certain outputs in response to certain inputs. For instance, photoreceptors and retinal ganglion cells are treated as black boxes, firing in response either to light exposure or other neuronal input. What black boxing accounts do *not* do is to account for the mechanistic underpinnings of a black box's input-output relationships. And though many explanations in biology (and elsewhere in the high-level sciences) are more mechanistic in flavor than the Hermann Grid account, they still usually appeal to black boxes of their own. For example, though a molecular biologist would rarely black box an entire cell, a protein or ribosome may be so treated. At the other extreme, some evolutionary and ecological explanations, which elide the mechanisms of mutation and the detailed workings of both an organism and its environment, exploit—explicitly or otherwise—black boxes at a grander scale.

Given the pervasiveness of black boxing, it is remarkable that even the most abstraction-friendly causal explanatory theories cannot recommend the accounts that feature them. For instance, as noted already, Woodward's (2003) influential view is that an explanation is explanatory in virtue of answering what-if-things-had-been-different questions (w-questions). The Hermann Grid explanation will answer *some* w-questions, such as where I would see illusory gray smudges were I

[3] See Kitcher (1999), Levy and Bechtel (2013) for a variety of alternative examples of abstract explanations from different biological domains.

to gaze at other printed patterns. However, an explanation that did not black box would answer strictly more w-questions, as it would be capable of accounting for what might have happened had the conditions required to maintain the black boxes' input-output relationships been disrupted. Thus, Woodward's proposal can make sense of the *minimal adequacy* of black-boxing explanations, but will invariably rule them *less* explanatory than any lower-level alternative.[4] Similarly, Strevens' kairetic account requires that mechanistic underpinnings be included in complete 'stand-alone' explanations. In consequence, the kairetic framework—while able to make sense of other forms of explanatory abstraction—never judges a causal model containing black boxes explanatorily optimal.

Why then, from a causal-explanatory point-of-view, are black boxing explanations so common in the sciences? *Pragmatic* considerations provide one tempting answer. Perhaps scientists present black boxing explanations in the form that they do because adding mechanistic details—though preferable on purely explanatory grounds—would overwhelm our limited human minds. Or perhaps each scientific subfield, for reasons of efficiency, provisionally treats high-level entities as explanatorily basic, even though complete explanations will ultimately banish black boxes entirely. Indeed, if you include enough 'pragmatic considerations,' you can almost always find a way of accounting for the explanations that scientists actually provide.

Yet, for those who take scientific practice as the key benchmark against which a theory of explanation should be judged, the failure of causal accounts to make room for explanatorily optimal black-boxing accounts should be distressing. Even granting that this form of abstraction can *sometimes* be accounted for on practical grounds—e.g., by the fact that we simply do not know what links certain inputs to certain outputs—its ubiquity in the explanatory annals suggests that there may be a distinct *explanatory* merit in these detail-sparse accounts. At the very least, the enormous gap between actual explanatory practice and the rulings of extant accounts seems reason enough to undertake the project of this paper: to pursue an explanatory theory that *prescribes*, rather than just *tolerates*, the kind of abstract explanations just reviewed.

[4] For discussion, see Weslake (2010: §2) and Franklin-Hall (2016a).

What problems must such a theory address, if it is to be successful? Here are two: the *stop problem* and the *carving problem*.[5] The *stop problem* is the challenge of articulating a vision of scientific explanation according to which optimal explanations should, on *explanatory* grounds, omit mention of mechanistic underpinnings even though they are causally relevant. After all, at least *some* of the happenings temporally between inputs and outputs of photoreceptors are difference-makers for output events and are part of the causal processes leading to them. So how could it ever be explanatorily preferable to omit their mention?[6]

The *carving problem* is the challenge of specifying just *which* black boxes should be appealed to in an adequate or optimal explanation. After all, for any target phenomenon there are numerous ways of representing the system responsible for it with schemes of interconnected black boxes, with each such scheme characterizing veridical and causal relationships. For instance, the Hermann Grid Illusion explanation above might have treated not *individual photoreceptors* as black boxes, but instead collections of them; equally, it might have treated as black-boxes entities that cross-cut neurons as they are customarily individuated. In consequence, an illuminating explanatory account—one that helped explain explanatory practice and did not merely describe its surface features—would need to offer principles that specified which black boxes were explanatorily legitimate, and which were unacceptable and gerrymandered.[7]

4—Causal Economy in Outline

In describing the Causal Economy account, I will focus on event explanation, which is arguably the most difficult case for an account of high-level explanation to handle. (I believe extensions to probabilistic, regularity, and contrastive explanations are relatively straightforward, but for reasons of space they must here be left aside.) For the sake of simplicity, I will further focus on a relatively simple kind of explanation—what I call a *direct event explanation*. Direct event explanations appeal to a state of affairs and at most one causal principle linking inputs and

[5] I articulate these challenges in my [2016b], where they used to evaluate mechanistic accounts of explanation.

[6] On the 'stop problem,' see Block's (1990) 'reductionist cruncher' and Jackson and Pettit (1992).

[7] A third important problem facing abstraction-friendly explanatory accounts, the 'disjunction' or 'overshooting' problem—pressed in my [2016a] against one version of Woodward's view—must be put aside here for reasons of space.

outputs. Some of these are, as it were, black-boxing accounts appealing to just one black box. Yet even direct explanations can either call upon a detailed state of affairs or high-level, abstract one, so this focus does permit me to display the basic capacity of the Causal Economy view to prefer high-level accounts.

The explanandum event might be perfectly concrete, individuated by all of its intrinsic properties (e.g., the fine-grained spilling of a glass of milk, with one droplet sinking into the carpet, another splattering on the cat's nose, etc.), or it might be individuated in a coarse-grained way (e.g., the mere emptying of the glass of its contents). The aim of explanation, on the Causal Economy account, is to identify the most important causal factors underlying some phenomenon, those factors in virtue of which the explanandum event was, to a rather substantial degree, *bound to happen*. The account's characterization of what constitutes 'the most important causal factors' has two parts: (1) a thin notion of causal influence and (2) a substantive 'selection principle,' which aims to isolate the explanatorily relevant aspects of the causal web.

The starting material from which an explanation is to be constructed is the basic causal tissue of our universe, which I will call *causal influence*. Any feature of the universe—whether a law or a state of affairs—in virtue of which any particular event happened *just as it did* is a causal influence on that event. Given the long reach of gravity and other physical laws, this means that causal influence is extensive; anything in the past light-cone of an event is a causal influence on that event. Thus, among the causal influences on my sneeze will be, not just the pollen in the air today and my hyper-active immune system, but also the wanderings of a particular wombat in the Australian Outback. Events, therefore, have countless causal influences extending far out in space and far back in time.

How should we understand the metaphysics of causal influence? Should it be understood in counterfactual terms, by appeal to regularities, the transfer of conserved quantities, or something else? Though some theories of causal explanation take a stand on such weighty matters, I will remain non-committal on the metaphysical nature of causal influence; it is possible to understand it in any of the ways just mentioned.[8] The one requirement I make is that causal influence is fully physical and thus describable in physical terms. It is best to understand causal influence physically, first, because of the appreciable evidence that our universe's movers-and-shakers are indeed physical laws acting on physical systems, and

[8] See Strevens (2008) for further discussion of this kind of approach.

second because a defense of high-level explanation is more powerful dialectically if it presumes a stark physical foundation.

Though *all* explanatory factors are causal influences on the Causal Economy view, not *all* causal influences are explanatory. Given that we want a theory faithful to actual explanatory practice, a complete event explanation will describe only a certain package of causal influences constituting the most important parts of the exceedingly complex complete causal story of an event. To pick out these *major causal contours* for an event is the task of the *selection principle*, which can be thought of as a sieve into which you feed packages of causal influence to separate out the packages of explanatory gold from the mounds of unimportant granite that surrounds it.

And just which causal influences should be selected as explanatory? If we want the major causal contours to correspond reasonably well with the content of actual explanatory texts, then we will want to include what they include and omit what they omit. As far as what should be omitted, this breaks down into three categories, informally described as follows:

(1) Causal influences with little consequence on the target event (like that of the antipodean wombat on my sneeze);
(2) Causal influences with a large effects on the target, but systematically treated as 'background factors' (like the role of oxygen in explaining a forest fire);
(3) Concrete details of certain causal influences when the specifics don't matter (like the particular location of particular transcription factors in a cell, rather than their overall concentration, in accounting for the cell's protein production).

That any single selection principle might successfully deal with all of these aspects of explanatory practice may sound too good to be true. After all, these issues have traditionally been approached separately. For instance, the second represents what is often called the 'causal selection problem,' and is usually disposed of in pragmatic terms. The third concerns part of the problem of proper 'explanatory level,' and while it is more often thought amenable to a principled solution, there is no consensus on what that is. I will suggest, however, that all

three of these aspects of explanatory selection spring from a common source and can be elucidated with a common selection principle.

In general, the core feature of the principle I recommend is that correct explanation is the product of *trade-offs* between two features, one that drives explanations to be more concrete and fine grained and another that rewards higher level of abstraction. In particular, an event is explained by a package of causal influences that maximizes the ratio of delivery to cost.[9] That is, the package will include those causal influences that make it the case that the target event was, to as great a degree as possible, bound to happen (thus 'delivering more'), while containing as little information as possible about the complete causal run-up to the event (thus 'costing less'). In this way, complete explanations give you the biggest *bang-for-your-buck*.

These notions will be elaborated in detail in the following section, but very briefly, a package costs less when it rules out fewer ways that the world might be. A package delivers more when, given that its causal influences took place, the explanandum event would still have taken place in spite of a greater number of circumstances having been different. As a rule, cheapness and delivery trade off: cheaper explanations deliver less; costly explanations deliver more. This can be illustrated at the extremes: *all* of an event's causal influences (which would be fully inventoried in an 'ideal explanatory text' on Railton's view) ensure that event's happening. At the other extreme, the thinnest sliver of the causal-influential nexus will offer very little stability: things could have gone just a bit differently, consistent with the influences cited in explanans, and the event would not have occurred. At the extremes, then, you get what you pay for.

Yet trade-offs can be more or less acute. In our universe there are 'sweet spots' where cost and delivery do not increase in lock-step, and it is here that the best explanations—those the Causal Economy account judges complete—are located. Complete explanations piece together the bits of causal influence that out-do themselves, offering disproportionate stability to the target event for their cost. As will be illustrated presently, these are found wherever there are robust processes, which are exceedingly common in biology and the high-level sciences more generally. It is largely because of such processes that high-level, detail-sparse explanation is possible at all.

[9] I leave open the possibility that there may be different 'currencies' in which these trade-offs might be formulated.

To take a toy example, consider an explanation for a ball's being located at the bottom of a bowl. In the lead up to this event, the ball was placed at a particular spot on the inside rim, which would be among the explanandum event's influences. To cite this *particular position* in explanation of the ball's final destination, however, is costly. A better explanation would specify that the ball was placed *somewhere within the rim*. This explanation is obviously much cheaper. And it delivers quite a lot: given that the ball was somewhere in the rim, lots of other things might have been different and the ball would still end up at the bottom. Another illustration comes from the explanation of a neuron's firing. One explanation might describe the neuron's complete physical architecture. But a cheaper one—typical of some high-level accounts that treat neurons as black boxes and only describe their input-output behaviors—would cite just that it was exposed to neurotransmitters above some concentration. Given this exposure, due to the robustness-making architecture of neurons (e.g., the large-scale redundancy of ion-based depolarization), lots of things might have been different and the firing would still take place. Thus it too is a cheap account offering rather substantial delivery.

Because there can be different robust processes—often at different spatial or temporal scales—involved in the lead-up to some events there might be multiple ways of achieving good economy. An explanation with low cost and modest delivery might be comparable to another explanation with higher cost and more substantial delivery, as they both offer the same *ratio* of delivery to cost. This is a welcome result if we want to account, in a principled, non-pragmatic way, for the diversity in explanatory practice that actually exists in and between scientific disciplines.

This tolerance for diversity notwithstanding, the selection principle will not be 'anything goes.' The principle has substance in virtue of ruling out many putative explanations. For instance, explanations that carve systems in 'unnatural' ways will be rejected, since these are usually descriptively costly without offering proportionate stability. Consider, for instance, an explanation for a neuron's firing—that is, its release of neurotransmitters at its axon terminal—that carved it into what we would normally consider a gerrymandered way. This release might be explained by citing the state of the cell body—rather than the axon terminal—at some prior time. Yet to do this in way that said enough to specify that that structure was in the midst of firing would be rather complex: it would need to at least detail the state of numerous ion channels and an ion differential across the membrane

surrounding the cell. This will be much more costly than one appealing only to the neuron's exposure to some high concentration of neurotransmitters. The selection principle also addresses the stop problem: even granting that low-level explanations do *deliver* more than abstract ones, they are also expensive. Thus it does not follow that fundamental-level explanations are always superior to more abstract, higher-level alternatives.

5—Causal Economy in Detail

I've just described Causal Economy view in a rather general way, to offer a flavor of the approach. This section will delve into the details of its three key concepts: *packages of causal influence*, *cost,* and *delivery*. These details are somewhat involved, but are important if we are to evaluate the account's capacity to describe the structure of actual scientific explanations without having simply presumed a high-level individuation of the universe into parts and kinds.

In doing this, however, I will not be concerned with another task that has received more attention from philosophers (though more commonly metaphysicians, rather than philosophers of science): showing that the account can reproduce judgments about the best explanation of events in systems that are not usually studied by scientists, such as in certain kinds of complex preemption scenarios. My view is that the principles governing our scientific explanatory practice—those that it is the job of the philosopher of explanation to uncover and evaluate—were trained on the kinds of causal systems scientists actually encounter. We do, I admit, sometimes have relatively strong intuitions about what causes (or causally explains) events even elsewhere, when events are said to be brought about by mechanisms designed to thwart causal-explanatory theories. It is an interesting question both why this is the case and what principles generate our judgments. But both because it has proven extremely difficult to capture these judgments in a unified way, and because it is hard enough to offer principles that honor the structure of explanations in the actual systems that scientists study, I must restrict my focus to these cases.

A—Packages of Causal Influence

The selection principle picks the major causal contours from a large set of *packages of causal influence*. But just what are these? They can be understood as having been created, in a 'production' step, from the complete causal story for an event. Given

the terrifically interconnected universe in which we live, the complete story includes, as noted above, everything in the past light-cone of the event: all states of affairs are influences, and all laws in virtue of which those states of affairs are influences are themselves influences. Assume a representational scheme expressive enough to describe all of this. (Given that causal influence is fundamental-level, this will be a fundamental-level language.) The many distinct packages of influence can be produced from it by either omitting mention of causal influences and/or by representing some of these influences more abstractly.

The idea of omitting certain causal influences is straightforward, but the idea of abstraction calls for some comment. Abstraction can work in three ways: by coarse-graining, amalgamation, and populational transformation. In *coarse-graining abstraction*, factors are not specified precisely, but only as falling within some range, or above or below some threshold. For example: 'over 15kg*m/s' rather than '20kg*m/s.' In *amalgamating abstraction*, the features cited are combinations of lower-level parameters. For example, 'momentum of 20kg*m/s' rather than 'mass of 2kg and velocity of 10 m/s.' In abstraction by *populational transformation*, the factors cited are population-level features, like temperature or concentration, rather than individual-level ones, like kinetic energy.

When these omission and abstraction procedures are applied—in every possible order and extent—to the representation of the complete story, a host of limited representations will result. Each of these rehearses just some part of the complete causal saga and constitutes a single *package of causal influence*. Some will include very little of the causal-influential story, some much more, some peculiar and disconnected bits. As an idealization, we imagine running each package through the selection principle, and have it measure modal delivery/descriptive cost. (In practice, of course, we never actually examine every possible causal package, but work from a highly compressed short-list.) Those packages that are maximal by this measure are the *major causal contours*, and constitute complete explanations. But just what is cost, and what is delivery? We turn to these next.

B—Cost

An explanation costs less in virtue of specifying fewer of the causal influences on an explanandum event and in less detail. To specify fewer causal influences and

in less detail is understood in terms of ruling out fewer ways the world might have been.[10]

It is easy to see that the omission of small influences, background factors, and features that are very concrete are all preferred when we favor less costly explanations. For example, an explanation for a particular neuron's firing would cost less by omitting mention of the small influence of a nearby massive object with only gravitational effects on the neuron. Likewise, an explanation that stated only that the neuron was at some prior time exposed to neurotransmitters would cost less than a rival explanation that also mentioned the background factor that sodium ions were above some specific concentration outside the cell wall. Finally, an explanation that cited the *concentration* of neurotransmitters to which the neuron was exposed would cost less than one that detailed the overly concrete description of the particular location and trajectory of each neurotransmitter molecule.

Yet cost considerations alone threaten to also omit the factors that *do* appear explanatorily relevant, since they cost something as well. This is where the delivery metric comes in.

C—Delivery

Defining Stability Boost

The delivery metric is the most complicated part of the Causal Economy picture. An explanation *delivers more* to the extent that it specifies a package of causal influence that provides a greater 'stability boost.' An event is *more stable* when the event occurs in spite of a greater number of things being different. A package of causal influences, therefore, boosts stability to the extent that the explanandum event is more stable when those causal influences occur than otherwise.[11]

The delivery or 'stability boost' of a package of causal influences can be understood to be the difference between two values, which I call the *construct stability* and the *baseline stability*. Put simply, the construct stability is the number of nearby possible worlds in which the explanandum event occurs, when fixing, in each of those worlds, that the causal influences specified in the explanans also occur. The

[10] To give substance to the idea of ruling out more or fewer ways the world might have been, we require a fine-grained and non-gerrymandered physical scheme of world individuation. I am assuming such a scheme is in principle available.

[11] Others who have emphasized the importance of stability in causal-explanatory contexts, with different conceptions of stability on offer, include Craver (2007), Lewis (1986: postscript C), Mitchell (2000), Woodward (2006).

baseline stability is the number of nearby possible worlds in which the explanandum event occurs *simpliciter*, that is, not fixing (in either direction) whether the causal influences are as specified in the explanans.[12] So:

stability boost = construct stability – baseline stability

Excepting cost considerations, complete explanations have a maximal stability boost. This reflects, I submit, the stability that the causal influences cited in the explanans provide to the target event.

So far I have helped myself to a set of nearby possible worlds. How should they be understood? Let us call the relevant set of nearby possible worlds the *close range*. This is a collection of worlds in which things went 'just a bit differently' than in the actual world, deviating from the actual world due to one or more small perturbations. The close range can be envisioned as a layered onion of possibility surrounding the actual world, where the inner-most layer contains worlds in which, by a Lewisian 'small miracle,' there has been a *single perturbation* on some component of the world, deflecting the course of that world from that of our own, howsoever slightly. That inner layer, in particular, contains a world corresponding to every possible single perturbation that could be carried out on the actual world. To produce the second layer, take each member of the first layer, and produce new set of worlds by executing an additional perturbation. This procedure is then iterated some specified number of times to produce the full close range.[13]

To define baseline stability from the close range, we begin by allowing all the worlds contained in it to unfold according to the physical laws until the explanandum event either definitely does or does not occur. The baseline stability is the number of worlds in the close range in which the explanandum event does occur.

Defining construct stability is more complicated. Recall that this is supposed to measure the number of nearby possible worlds in which, having ensured that the

[12] As will come out below, because the baseline stability for a target event is the same for all candidate explanans, the work of the measure is done by the construct stability. Yet it is still notionally helpful to see the stability boost as the difference between these factors.
[13] Several of the key parameters of this characterization have been left vague. These are discussed in more detail below. These specifics are postponed partly to avoid cluttering the exposition and partly because there may be different ways of filling out those details and the big picture is what matters more.

causal influences specified in the explanans do occur, the explanandum event also occurs. To define construct stability, we allow the worlds in the close range to unfold according to the laws until the causal influences cited in the explanans have definitely occurred or not. In those worlds in which the causal influences do not occur, we introduce them by a small miracle. We now have, not our initial close range of possible worlds, but a new construct that is sibling to it in which the relevant causal influences occur in every world. Next we let these worlds unfold according to the laws until the explanandum event definitely does or does not occur. The construct stability is the number of worlds in this sibling range in which the explanandum event occurs.

As I've already noted, the stability boost is the difference between construct stability and baseline stability. It represents the *additional stability* that the candidate explanans provides the explanandum.

D—The Parameters of the Close Range

With this definition of stability boost in hand, it is worth saying something about the parameters that went into defining the close range: the *components* of the world are candidates for perturbation, what a *single perturbation* is, the *time* at which the worlds of the close range are first permitted to diverge, and the *number* of perturbations in the outermost layer of the close range.

What are the basic components of the world that are candidates for perturbation? In principle, these might be restricted to components within a contained system (e.g., the cell, the visual system, etc.) or they might range over the whole universe. Further, these components might be composite or 'high-level' individuals (like populations or cells) or more basic ones (like fundamental particles). I will opt for the latter, low-level components only, and will permit perturbations on any feature in the world, not just those constrained to a particular system. In this way, I will not have to presume a controversial high-level scheme of individuation of either parts or of wholes, which I see as the output of an explanatory theory rather than its input.

What counts as a *single* perturbation will partly depend on how we have divided the world into basic components, since the perturbation must affect some feature of those components. For example, if we divided the world into atoms, characterized by their physical location and electronic configurations, then a single perturbation might affect either of these properties of a single atom in a small way.

But what do we mean by a *small way*? For instance, how far can the atom be moved, or its electron's state elevated? Given that we are working with a fine-grained division of the world, I believe it does not matter much how we answer that question. This is because most of the action in distinguishing some packages of causal influences from others is going to take place some considerable distance from the actual world, in layers containing worlds into which a great many perturbations have been introduced. These more substantial changes could have been constructed from very different ways of defining a single perturbation.

At what point in *time* in the actual world do we start introducing perturbations? In principle, it might be any time before the explanandum event, and (given the way we measure delivery) our choice of this parameter will determine the earliest time at which a causal influence can be even potentially judged as providing any stability boost. The choice of the time parameter is therefore quite consequential. But, unless one goes back to the beginning of time, it seems that it must also be conventional. This conventionality, however, should not surprise us. In searching for explanations, scientists do tend to narrow their temporal gaze.

The final parameter in the setting the close range is the number of iterated perturbations, which sets the maximal difference between the actual world and any particular world in the close range.[14] This must be conventional as well, but not in any threatening way. Analogous to the way the time of perturbations reflects the 'temporal gaze' of our explanatory practice, this feature is indicative of our 'modal gaze.'

6—Causal Economy in Practice

Causal Economy's emphasis on an explanation's *costing less* and *delivering more* work together by pulling in different directions. Cost considerations drive explanations toward more abstraction, but it would be miserly madness to value cost alone as if the 'empty explanation' were the most satisfying. Considerations of delivery, by comparison, push explanations toward the more concrete and detailed in a bid to increase the stability of the explanandum event. But, if price were no object, then we would end up with impossibly extravagant explanations that would invariably detail the workings of vast tracts of the universe. So, in order to account

[14] I take this number to be finite, since there are only a finite number of ways that the world might be. The measure I offer, in turn, depends only on a subset of that finite number of possibilities.

for the kinds of explanations that scientists actually provide, Causal Economy plays cost and delivery against one another. How, then, will Causal Economy work in practice? That is, restricting our attention to direct event explanations, what sorts of causal influences will be included and excluded in explanations?

Recall the way we defined delivery in terms of stability boost, and stability boost in terms of two values: baseline stability and construct stability. We determined baseline stability by counting the number of possible worlds in the close range in which the explanandum event occurs. Then, to determine construct stability, we introduced the causal influences cited in the explanans in every possible world in which they do not already occur as a matter of course, and again counted the number of worlds in which the explanandum event occurs. Finally, we subtracted the number representing baseline stability from the number representing construct stability. This means that the value representing stability boost will be equivalent to the number of worlds in which we introduced the relevant causal influences and the target event occurred.

The upshot of this is that stability boost is higher for a causal influence when it occurs as a matter of course in fewer worlds in the close range of worlds *and* when the explanandum event occurs in many worlds in which we have introduced the relevant causal influences. These will, in general, be causal influences that (a) are themselves rather unstable (i.e., were they taken as a target event their baseline stability would be small), but (b) which are difference-makers for the target event. Why should my algorithm select these factors? Speaking to (a), only if a causal influence is itself unstable might it be disturbed by the perturbations that distinguish other worlds in the close range from our own. This is required if the miracle that reinstitutes that influence is have any consequence. Speaking to (b), only when a causal factor is necessary for the occurrence of the event will this reinstituting miracle even potentially eventuate in the occurrence of the explanandum event in worlds in the close range.

Consider again the explanation of the retinal ganglion cell's release of neurotransmitters in our example of the Hermann Grid Illusion. This event is the last step in an extended process of neuronal firing and has countless causal influences, including (to mention only those close at hand) a suitably high ATP concentration, the correct partition of ions across the cell membrane, and various properties of the embedded membrane channels. Of the myriad causal influences, the one that will *invariably* be cited in any explanation concerns whatever 'activated'

the neuron—in this case, exposure to neurotransmitters by another neuron. This itself is a relatively high-level or coarse-grained event and is not described in all its details. And although this will never be causally sufficient to bring about the target event, it is the one that best satisfies the above conditions (a) and (b). Exposure to neurotransmitters is not itself a particularly stable event, since the relevant neuron would not have been exposed had other neurons not fired as they did, but given that exposure the neuron would have fired even had many other things been different.

Someone might object that this kind of account is going to cite the wrong causes as explanatory in certain kinds of pre-emption scenario. We can imagine cases in which the 'actual cause' would not have brought about the target event had conditions been even slightly different, but the target event would still have taken place due to 'back-up' factors that would be effective in a wide range of circumstances. Causal Economy will pick out the back-up factors as more explanatory than the actual cause, since they provide greater stability boost, and that might intuitively seem wrong. There are several things to say in response. First, in the most common kind of pre-emption case, one in which the back-up factors are similar to the actual cause, it is a virtue of the Causal Economy account that it includes the back-up factors in the explanation by abstracting away from particular causes. Thus, biologists would not explain a neuron's release of neurotransmitters in terms of the very neurotransmitters it was actually exposed to, but in terms of the concentration of neurotransmitters released in the vicinity. In this way, the back-up factors and actual causes are bundled together as part of the same explanatory package.[15] But we can cook up examples where the back-up factors are very different than the actual causes. One thing to say about these cases is that, while common in philosophical discussions of the metaphysics of causation, they are rather rare in actual scientific contexts. It should come as no surprise if the norms implicit in scientific practice are suited to routine cases, not outliers. We can also observe that pre-emption is a thorn in the side for many theories of causation and explanation (especially those employing counterfactuals). Accounts that are specially designed to deal with pre-emption generally pay the price of being less satisfactory in accounting for explanations more central to actual scientific practice.

[15] See Nathan (this volume) for an alternative take on such back-up factors. In particular, on Nathan's account, backups are packaged into the *cause* rather than the *explanation* for an event.

If Causal Economy does very well in accounting for the sorts of explanations that are central to scientific practice, we have reason to run with it and see how far we can take it. Perhaps with suitable ingenuity we could work out some conditions to deal with pre-emption puzzles, but this is not the place to delve into these Gettier-like complications.

Having said this much about which influences Causal Economy picks out as explanatory, what sorts of causal influences will be left aside? The answer is, corresponding to the two above conditions, those that are highly stable themselves, and those that do very little to make the explanandum event more stable, not even being difference-makers for it. Let us take these in reverse order. Although we would include every detail that had any effect on the target event *if* we were concerned with delivery alone, Causal Economy will regard causal influences explanatorily unimportant when they do little to make the target event more stable in comparison to their explanatory cost. Thus, we will entirely omit causal influences that make very little difference to the target event and leave out concrete details that add little explanatory power to a more abstract characterization.

Causal Economy will also judge unimportant causal influences that are themselves highly stable. While being strictly necessary for the occurrence of the explanandum event, these causal influences are themselves so stable that they occur as a matter of course in virtually all worlds in the close range—and consequently, there is little or no scope for them to boost the stability of the explanandum event over its baseline stability. This might seem odd at first. If the relevant causal influences occur as a matter of course in many nearby possible worlds, then why shouldn't these be considered explanatory if they occur together with the explanandum event? The answer is that these highly stable causal influences are what we usually consider background factors. It is one of the chief virtues of Causal Economy that it provides us with a principled reason to exclude these background factors from our explanations, which seems consistent with our actual explanatory practice—though this is often judged a shortcoming by philosophers with more a priori standards for what constitutes a complete explanation.[16]

Saying that good explanations should maximize the ratio of delivery to cost assumes some way of weighing these two values against one another, since they are not expressed in a 'common currency.' What is the correct way of balancing them? I won't try to offer an answer. After all, this is a hard question endemic to all

[16] I further develop the Causal Economy treatment of background factors in my [2015].

'economic' problems involving a trade-off of one good against another. Economists often assume that the weights assigned to different goods are pure subjective preferences. We might strive to discover a less subjective weighting through an examination of the actual weightings employed in scientific practice (a possible research program in experimental philosophy of science). Though it is plausible to think that there is some range of delivery-to-cost trade-offs that scientists do typically employ, it may well be that these weightings are not *fully* objective, and do depend on facts about our cognitive capacities and interests. This lack of full objectivity, however, is not as alarming as it may first appear. For even if it is something about us that determines the weightings, the very existence of 'sweet spots' where we get more explanatory bang-for-our-buck is a feature of the causal architecture of the world.

Causal Economy can, thus, account for all three categories of things that we have observed are usually left out of actual scientific explanations: (1) causal influences with very little effect on the explanandum event; (2) causal influences that may have a significant causal influence, but are so stable that they are usually regarded as background factors; and (3) insignificant concrete details of important causal influences.

In describing how Causal Economy works, I have for the sake of simplicity been focusing on direct event explanations—explanations that cite a particular state of affairs and at most one causal principle linking inputs and outputs. These are, as I have said, essentially explanations appealing to at most one black box. More complex black-boxing explanations, like the full explanation of the Hermann Grid Illusion, describe a system of interconnected black boxes. But the way Causal Economy handles these is not radically different from direct event explanations. In particular, we simply view these complex black-boxing explanations as a collection of direct event explanations.[17] In the simplest case, we have two black boxes, one

[17] This requires one qualification. There are, to be more precise, two kinds of black boxes in actual explanatory practice. A *physical black box* is one in which inputs and outputs of each black box are described in physical terms, however abstract. A *functional black box* is one in which either the inputs or outputs are not described in physical terms, but rather in terms of their functional relationships with other black boxes. The example of the Hermann Grid Illusion featured some functional black boxes. In particular, instead of describing the output of photoreceptors physically (for instance, by describing this as a 'neurotransmitter release'), it was described in terms of its effects on retinal ganglion cells (they were either activated or inhibited). As I see it, functional black-boxing explanations leave out the physical links between boxes for expository convenience. This is indeed a

of whose output is the other's input. We can think of this as two direct event explanations, defined by the input and outputs that were actually realized in a particular case. In deciding between rival explanations, we compare the causal economy of each set of direct event explanations. More needs to be said about exactly how these comparisons will be made, in particular about how we should compare the economy of direct event explanations of separate events. But the general principle of evaluation will remain the same: ask whether that additional detail required of one scheme of black boxes boosts the stability of the target event enough to justify the additional cost.

7—Rationalizing the Causal Economy Standard

I have suggested that explanatory trade-offs—in particular those between abstractness and stability-boosting—have the potential to account for the kinds of explanations that are actually offered by the high-level sciences. Alternative theories of explanation must instead appeal extensively to practical considerations to bridge the gap between real and ideal explanations. But someone might object that Causal Economy account avoids doing the same only by baking pragmatic considerations into the selection principle. After all, why should explanations—other things equal—*be more abstract* if it weren't simply because we happen to have limited minds unable to process in full the complex causal nexus?

My first response is to recall the basic motivation for the Causal Economy view: to identify an event's most important causal factors—those for which the event was, to a substantial degree, bound to happen. These factors reflect, as I've noted above, 'sweet spots' in the architecture of our universe. In this way, the demand that explanations be abstract was not an end in itself, but instead a vehicle for picking out what seems intuitively to be the positively explanatory features.

One might wonder whether this reply leaves the original question unanswered, in a slightly different form. Why are just *these* factors explanatory? Why shouldn't complete explanations cite *more*—all of an event's difference-makers, or all its causal influences, or perhaps all its causal influences and metaphysical grounds (whatever those might be)?

---

pragmatic element of actual black-boxing explanations, albeit a limited one. If this is right, then functional black-boxing explanations are ultimately explanatorily parasitic on their physical counterparts that spell out the relevant physical properties. Therefore, to apply the Causal Economy account to a functional black box, one would first have to 'translate' it into a physical black box.

It is difficult to offer a satisfying reply to this line of questioning, but here the Causal Economy account is in the same situation as any other explanatory theory. Whatever story might be told about which factors are, or are not, explanatorily relevant, it is always possible to ask: why not more, less, or something different? The deep source of this stalemate springs from the unique nature of explanatory norms in the scientific project, which we can best appreciate by contrasting them with norms of inference.

Though there are arguments about the proper inferential rules, the common ground in debates of their merits is the fact that inference aims at true belief. Therefore, inference principles can be evaluated based on their capacity to deliver this result. At what does explanation aim? Presumably, full understanding. While we have something like an independent grasp on truth, our only access point to what full understanding consists in is our notion of what constitutes a correct or complete explanation. Thus it is hard to stake out neutral territory from which we might debate the relative merits of one or another picture of explanation or understanding. In light of this difficulty, we can only demand of an explanatory account some conception of what is explanatory that 'rings true.' If we cannot place strong constraints on the explanatory principles on these a priori grounds, what should constrain our explanatory theorizing? As I've already emphasized, I take explanatory practice to be my guide. Best for actual scientific explanations to be judged correct or optimal as they stand. Given that this essay has focused on applying the Causal Economy account to just a particular kind of explanation—and has dwelled on just a few examples—it remains to be seen how well its trade-off metric captures the structure of the many kinds of explanations in scientific circulation. But it seems at least a promising way to begin thinking about the special virtues of explanation in the higher-level sciences.

References

Andersen, HK (2017), 'Pattens, Information, and Causation', *The Journal of Philosophy,* 114 (11), 592-622.

Appel, B., Givan, L.A., and Eisen, J.S. (2001), 'Delta-Notch signaling and lateral inhibition in zebrafish spinal cord development', *BMC Developmental Biology,* 1 (1), 13.

Baumgartner, G. (1960), 'Indirekte Größenbestimmung der rezeptiven Felder der Retina beim Menschen mittels der Hermannschen Gittertäuschung', *Pflügers Archiv European Journal of Physiology,* 272 (1), 21-22.

Block, N. (1990), 'The computer model of the mind', (MIT Press), 247-89.

Franklin-Hall, LR (2015), 'Explaining causal selection with explanatory causal economy: Biology and beyond', *Explanation in biology* (Springer), 413-38.
--- (2016a), 'High-Level Explanation and the Interventionist's' Variables Problem", *British Journal for the Philosophy of Science,* 67 (2).
--- (2016b), 'New mechanistic explanation and the need for explanatory constraints', *Scientific composition and metaphysical ground* (Springer), 41-74.
Hitchcock, Chris and Woodward, Jim (2003), 'Explanatory generalizations, part II: Plumbing explanatory depth', *Noûs,* 37 (2), 181-99.
Kitcher, P. (1981), 'Explanatory unification', *Philosophy of Science*, 507-31.
--- (1984), '1953 and all that. A Tale of Two Sciences', *The Philosophical Review,* 93 (3), 335-73.
--- (1999), 'The hegemony of molecular biology', *Biology and Philosophy,* 14 (2), 195-210.
Levins, R. (1966), 'The strategy of model building in population biology', *American Scientist,* 54 (4), 421-31.
Lewis, D. (1986), 'Causal explanation', *Philosophical papers,* 2, 214-40.
Nathan, M.J. (2012), 'The Varieties of Molecular Explanation*', *Philosophy of Science,* 79 (2), 233-54.
Nathan, Marco (forthcoming), 'Redundant Causality and Robustness', in Ken and Wooward Waters, Jim (ed.), *Minnesota Studies in the Philosophy of Science*.
Palmer, S.E. (1999), *Vision science: Photons to phenomenology* (1: MIT press Cambridge, MA).
Pettit, Frank Jackson and Philip (1992), 'In Defense of Explanatory Ecumenism', *Economics and Philosophy,* 8, 1-21.
Putnam, Hilary (1975), 'Philosophy and our Mental Life', *Mind, Language and Reality* (London: Cambridge University Press).
Salmon, Wesley (1984), *Scientific Explanation and the Causal Structure of the World* (Princeton, NJ: Princeton University Press).
Sekuler, R. and Blake, R. (1994), 'Perception, 3rd', (New York: McGraw-Hill).
Snowden, R., Thompson, P., and Troscianko, T. (2006), 'Basic vision: an introduction to visual perception', *Recherche,* 67, 02.
Sober, E. (2000), *Philosophy of biology* (Westview Pr).
Strevens, Michael (2008), *Depth* (Cambridge, MA: Harvard University Press).
Weslake, B. (2010), 'Explanatory Depth*', *Philosophy of Science,* 77 (2), 273-94.
Woodward, Jim (2003), *Making Things Happen: A Theory of Causal Explanation* (Oxford: Oxford University Press).
--- (2006), 'Sensitive and Insensitive Causation', *The Philosophical Review,* 11 (1), 1-50.