Comments on Michael Strevens's Depth

Comments on Michael Strevens's *Depth*

Ned Hall, Harvard University

This first-rate treatment of explanation should help set the agenda for work in this area for years to come.

Now to the inevitably more critical points. I will zero in on two concepts central to Strevens's project: "causal entailment" and "cohesion". I will try to raise problems for the first, and then suggest how they might be solved, by taking an approach different from his own to the second.

Consider a well-known example:

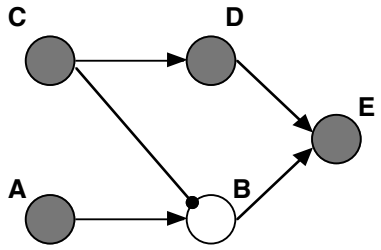

Figure 1

Circles represent "neurons"; arrows represent stimulatory channels; lines ending with dots represent inhibitory channels. Shading a circle indicates that the neuron fires. Temporal order is represented by reading from left to right. Bold capitals name neurons, italicized capitals events of their firing. Here, neurons **A** and **C** fire simultaneously (at time 0, say); **C** sends a stimulatory signal to **D**, causing it to fire (at time 1), while **A** sends a stimulatory signal to **B**. Since **C** also sends an inhibitory signal to **B**, **B** does not fire. Finally, **D** sends a stimulatory signal to **E**, causing it to fire (at time 2). Figure 2 shows what would have happened if **C** had not fired:

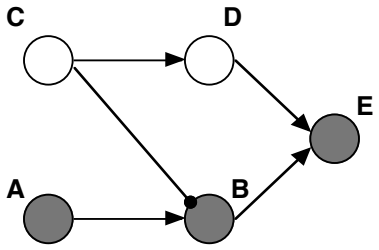

Figure 2

This example is a nice test case for a philosophical account of explanation: an account ought to show that, and why, *C* is explanatorily relevant to *E*, whereas *A* is not.

Suppose you favor a *causal* approach to explanation. As Strevens astutely explains, you face a choice. One path: you lean on some account of the metaphysics of causation, which you hope will demonstrate that *C* but not *A* is among the *causes* of *E*; you may then simply draw upon this result, saying that *C* is explanatorily relevant to *E* because it is among *E*'s *causes*, whereas *A* is not because it is not. Strevens abjures this path, preferring an alternative that assumes little about the *metaphysics* of causation. The aim is then to develop a sophisticated story about how causal information gets *packaged* into explanatory information – which story, if all goes well, will provide the tools for distinguishing the explanatory relevance of *A* and *C* in cases like this.

It is not that *no* purely metaphysical account of causation gets presupposed. The picture is rather this: The theory of explanation needs, as one of its raw ingredients, *some* such account. (Strevens considers several options, remaining neutral among them.) But this account can be *highly non-discriminating* – so much so, perhaps, that it qualifies *both A* and *C* as causes of *E*. (In Strevens's preferred terminology, both *A* and *C* may count as exerting *causal influence* on *E*.) Only after substantial processing and refinement that adds in certain other ingredients do we cook up the kind of sophisticated causal-explanatory concept we are deploying when we confidently assert that in figure 1, *E* fires *because of C*, and *not* because of *A*.

Let's plump for a specific account of the metaphysics of causation. (It's one of the accounts Strevens treats as a live option.) We first need Strevens's distinction between "concrete" and "high-level" events, which I will simplify as follows: take a "concrete" event to consist simply in the instantiation by some space-time region of some *complete physical state*. Stipulate that for any "high-level" event *C*, there is a set T of concrete events such that *C* occurs iff some member of T occurs. So concrete event *C** is a *realizer* of high-level event *C* iff the claim that *C** occurs entails the claim that *C* occurs.

Now for the account of causal influence: *C** causally influences *E** just in case, for a suitably wide range of alternative concrete events to *C**, had any one of those alternatives occurred in place of *C**, some corresponding alternative to *E** would have occurred in place of *E**. Taking for granted that the alternatives to *C** occur in the same region of space-time R_1 that *C** occupies, and similarly for *E**'s region R_2, here's another formulation: the contents of R_2 counterfactually co-vary to a suitably high degree with the contents of R_1. We can leave it open what "suitably" comes to. A natural approach is that *any* amount of counterfactual covariation secures *some*

causal influence of *C** on *E**, and that we measure the *amount* of influence by the extent of the covariation. These details will not matter.

Where *A** is a concrete realizer of *A* in figure 1, and *E** a concrete realizer of *E*, our diagram leaves it unspecified whether *A** causally influences *E**. But it's obvious that we can *stipulate* that this is so, without imperiling the intuitive verdict that *C* and not *A* causally explains *E*. Suppose that firing neurons emit radiation that has a substantial effect, not on *whether* any other neuron fires, but on *how* they do so, if they fire at all. Thanks to **A**'s firing, **E** fires in the color green; if **A** had not fired at all, the firing of **E** would have been red. Then not only does *A** causally influence *E**, but by any reasonable way of measuring *degree* of influence, its degree of influence will be considerable. The example thus shares crucial features with Strevens's Rasputin example, in which Rasputin dies by drowning only after the efforts of his assassins to poison him and shoot him fail to kill him. Strevens (p. 46) correctly uses this example to establish that relations of causal influence are too crude to provide a basis for drawing the explanatory distinctions we need.

So how *do* we show that in figure 1, *C*, and not *A*, is explanatorily relevant to *E*? By showing that *C* (more exactly: the claim that *C* occurs) is an essential part of at least one veridical causal model for *E*, whereas *A* is not.

I do not think this can be done.

A veridical model for some claim (e.g., the claim that *E* occurs) is a set of true sentences that entail that claim. Statement P is an essential part of some veridical model for statement Q iff the statements in the model aside from P fail to collectively entail Q. The trouble comes when we try to say what a *causal* model is. Here is Strevens's first explanation:

> Finally, the model is causal because the statements in the model do not merely entail that *e* occurs; they *causally entail e*, meaning that the entailment of *e*, or more exactly the derivation of *e*, mirrors a part of the causal process by which *e* was produced. (pp. 71-2)

This is no official definition; that comes later. Still, warning flags go up. There just is no such thing as *the* derivation of some claim from some other claims. Nor is it clear what sort of independent grip we're meant to have on the notion of "causal process", let alone the notion of *the* causal process by which some event was produced. Unfortunately, a look at the gory details doesn't seem to help.

A is certainly an essential part of a *veridical model* for *E*[1]: let this model consist of sentences P_1 – P_4 stating that (i) **A** fires at time 0; (ii) the five neurons are connected

[1] I'm sloppily eliding the difference between *A*, and a sentence stating that *A* occurs.

as shown; (iii) nothing outside the five-neuron network is happening that could interfere with how events within this network unfold; (iv) appropriate dynamical 'neuron laws' connect the signals received by a neuron to its behavior, and determine how signals are transmitted along channels. P_1 - P_4 will entail that **E** fires at time 2, but P_2 - P_4 by themselves clearly will not.

Is this veridical model a *causal* model – and, if not, for reasons that do not *also* spell trouble for the explanatory standing of *C*?

Set aside two confused grounds for doubt. You might worry about the propriety of some or all of P_1 - P_4. Don't: any decent veridical causal model that shows that *C* is explanatorily relevant to *E* will need P_2 - P_4, together with an exact analog of P_1. You might worry that "the" derivation of *E* from this model won't mirror the causal process by which *E* in fact comes about: don't we show that **E** must fire by using a disjunctive syllogism, reasoning that if **C** fires, then (for such-and-such reasons) **E** will fire, and likewise that if **C** doesn't fire, then (for such-and-such *different* reasons) **E** will still fire? Well, we *could* produce such a derivation. But this observation is doubly irrelevant: we don't *need* to; and at any rate, what could possibly entitle us to assert, at *this* stage of the analysis, that such a derivation *doesn't* mirror the causal process by which *E* is brought about? If we'd already produced an account of "causal processes" sufficient to secure this verdict, we could go home.

So let's examine whether Strevens's official account can rule out our model as non-causal. Here is one part of that account (extracted from p. 78): *every way of making the premises concrete causally entails some way of making the conclusion concrete*.

Does our model have this feature?

Suppose not. That will be for one of two reasons.

First reason: Some way of making the premises concrete fails to *entail* ("causally" or otherwise) any way of making the conclusion concrete. But if *that* is what stands in the way of finding a veridical causal model that witnesses the explanatory relevance of *A* to *E*, then – given the symmetry in the relevant entailment relations – it is very hard to see why we won't run into exactly the *same* trouble, when trying to find a veridical causal model that witnesses the explanatory relevance of *C* to *E*. So let's set this worry aside.

Second reason: Some way of making the premises concrete entails a way of making the conclusion concrete, but fails to do so *causally*. But how would this happen, exactly? The way in question will replace our P_1 with a premise stating that the state of the region R_1 taken up by *A* is S_1. It will replace P_2 and P_3 with a specification (perhaps partial) of the state of the *rest* of the world at the given time (say, that this state is Z). It will replace our 'neuron laws' with fundamental physical laws, or – better – with special-purpose consequences of them, that explicitly assert

that if R_1 is in state S_1 at some time, and the rest of the world is in state Z at that time, then 2 units of time later R_2 will be in state S_2 – where this will count as a concrete realizer of *E*.

Now look at what Strevens writes, in the preceding pages:

> Let me begin at the bottom, with the facts about causal entailment between concrete events.
>
> …
>
> A causal influence relation between events *c* and *e* exists just in case there exists a causal law connecting a property *P* of *c* and a property *Q* of *e*, a law having roughly the form *If P(c) and Z then Q(e)*, for some set of background conditions *Z*.
>
> …
>
> I now define causal entailment in the obvious way: the derivation of the *Q*-hood of *e* from the *P*-hood of *c* is a causal entailment just in case it goes by way of modus ponens applied to a causal law (or consists of a chain of such deductions). (pp. 76-77)

The trouble is that the model we just detailed meets these requirements to the letter.

Or does it? Maybe the laws included in this model – the laws that assert that if R_1 is in state S_1 at some time, and the rest of the world is in state Z at that time, then 2 units of time later R_2 will be in state S_2 – are not really *causal* laws:

> What is a causal law? … it is a logical consequence of the fundamental laws that satisfies a further condition whose content is dictated by the correct metaphysics of causal influence. … on the counterfactual view, the causal law must hold because, in the presence of *Z*, *e*'s *Q*-hood counterfactually depends on *c*'s *P*-hood. (p. 77)

I cannot make full sense of this condition[2], but also cannot see how our laws fail to meet it. (Remember that we've already stipulated that the actual and possible concrete realizers of *A* and *C* are *alike* in exerting substantial causal influence on the corresponding realizers of *E*.)

This is the end of the line: the search for grounds to deny that *A* is an essential part of some veridical causal model for *E* have turned up nothing remotely conclusive. No great surprise, I think: if we help ourselves to little more than entailment relations mediated by laws, we just cannot gain sufficient leverage to prise *A* and *C* apart.

So we should look elsewhere. Where? An attractive option comes into view, if we consider Strevens' important concept of "cohesion", and how best to understand it.

---

[2] Since it has already been stipulated that a causal law is a *logical consequence* of the fundamental laws, what does it mean to *add* that it holds "because" some relation of counterfactual dependence holds?

Begin with Strevens's extremely important insight about *explanatory depth*, and the way that one causal explanation can achieve *more* of it than another. Strevens ties this insight to his preferred account of explanations as provided by veridical causal models, but this is not necessary. Let's just assume that we're working with *some* approach to explanation that sees the explanation of singular events, at least, as consisting in the provision of certain kinds of information about their causes. Now consider a simple example:

A window has broken. Why? Because Suzy threw a rock at it. We could fill out that answer in many ways: we could trace the intermediate causes connecting Suzy's throw to the breaking; we could trace her throw's own causal origins; we could highlight the other causes contemporaneous with her throw with which it conspired in order to bring about the breaking. But Strevens brings out a distinct dimension along which our explanation of the breaking can be deepened: we could highlight those aspects of Suzy's throw *in virtue of which it was able to bring about the breaking*, distinguishing them from other aspects that were explanatorily *irrelevant*. For example, the mass of her rock was important, but its color, not so much. And we could go further still, articulating, even with some mathematical precision, the *structure* of the way in which the window's breaking depended upon such factors as the rock's mass, the angle and velocity of the throw, and the distance between Suzy and the window.

How do we capture, in general, this sort of depth-producing information? For Strevens, we do so by starting with a highly specific veridical causal model of our *explanandum*, and abstracting away details by means of a certain procedure. Myself, I prefer to think about this issue counterfactually: very roughly, we achieve depth by showing how to distinguish those counterfactual variations on Suzy's throw that would and wouldn't still lead to the window's breaking. But this difference of emphasis is minor. The key point is that, one way or another, we show how to situate the details of what *actually* happens within a suitably chosen range of *possible alternatives*.

The qualifier "suitably chosen" is essential. It's helpful – *deepening* – to note that the window would still have broken had the rock been a different color, and not if it had been sufficiently less massive. It's *not* helpful to note that the window still would have broken if, instead of throwing a rock, Suzy had set off a bomb. For that matter, it's not helpful to note that it still would have shattered if she had thrown so hard that the rock, while missing the window, broke the sound barrier, with the resulting sonic boom doing the work. Somehow, we are able distinguish *relevant* from *irrelevant* alternatives within which to situate the actual scenario. Strevens: what distinguishes

these relevant alternatives *as such* as that, taken together, they form a *cohesive* set. The question is how to understand "cohesion".

One preliminary point: the right notion of "cohesion" needs to operate *at the level of causal processes themselves*, and not just at the level of their "inputs". Consider two scenarios, each of which begins with the launching of a projectile and ends with a window shattering. The scenarios differ only slightly in the manner of the launch. But they differ just enough so that in one case, the window shatters because the projectile strikes it; whereas in the other case, it shatters because the projectile breaks the sound barrier. These should count as substantially *different* ways of bringing about a shattering. But you won't see that, if you look for the differences just in the initial conditions.

I follow Strevens in thinking that we are operating with some crucial notion of *similarity* here. Go back to the example of Suzy breaking the window. We collect together the actual scenario with alternative scenarios in which the window shatters as a result of *closely similar* processes. We achieve explanatory depth, concerning the question why the window shatters, by clearly articulating the features that all of these scenarios share in common, that serve to distinguish them from alternatives in which the window does *not* shatter. So the philosophical task is now to explain the nature and origin of the needed standards of similarity.

The only alternative Strevens takes seriously also strikes me as hopelessly optimistic: "Differences between causal elements are perhaps to be discerned at the level of fundamental physics." (p. 104) Where else to look? Not to metaphysics: I agree with Strevens about that. But another option can be discerned in the "unificationist" strand in our concept of understanding. The key idea is this: *part* of what we are after in explanation is the acquisition of *a cognitively effective means for organizing our information about the world.*

This desideratum is most visible in mathematics. In lieu of a proper theory, let me offer a somewhat silly but nicely illustrative example.

Consider the following initial segment of an infinite sequence of natural numbers:

1,1,1,2,3,2,1,3,5,4,2,5,7,8,3,7,9,16,5,11,11,32,8,13,13,64,13,17,…

Perhaps you find this sequence confusing. You don't *understand* it. You don't know *why it has the form it does*. If so, the following way of *reorganizing* the initial segment will make things crystal clear:

1, 1, 1, 2,
3, 2, 1, 3,
5, 4, 2, 5,
7, 8, 3, 7,

9, 16, 5, 11,
11, 32, 8, 13,
13, 64, 13, 17,…

Looking down the columns, we four simple, familiar sequences. Once you see this, you *understand* the original sequence. But *not* by acquiring a special sort of information about it.

We can't seriously doubt that this desideratum operates in empirical inquiry, as well. That leads to a suggestion: perhaps the "similarity metrics" by which we organize our view of causal processes in a given domain – and thereby come to see certain sets of them as more *cohesive* than others – earn their privileged status because of the way they contribute to the effective *organization* of causal knowledge. In a tiny bit more detail: What scheme for taxonomizing causal processes, and capturing relations of similarity among them, best lends itself to the articulation of a set of widely-instantiated, informative, and unified causal generalizations in a given domain? *That's* the scheme that will ground the judgments of similarity that in turn underwrite the standards of "explanatory depth" that Strevens has drawn attention to. If this approach to "cohesion" is correct, then a key piece of unfinished business is to construct a proper theory of the effective organization of causal information.

Suppose we have such a theory in hand. In light of it, we can see that, in a certain domain, a certain way of taxonomizing causal processes into types, and imposing a similarity metric on these processes, is *best*, given the aim of effective organization of causal information. Then a solution to the stubborn problems posed by cases of preemption (e.g., figure 1) may drop out. Why does *C*, in figure 1, count as explanatorily relevant to *E*, whereas *A* does not? Because there is a sequence of events connecting *C* to *E* that counts not *merely* as a chain of causal influence – such can easily be found, connecting *A* to *E*, as well – but counts as a process *closely similar* to a wide range of other such processes, both actual and possible, connecting firings of neurons to one another. If so, then the proper order of development of a philosophical theory of explanation is not quite the one Strevens follows: while we should agree that the ultimate *metaphysical* ingredients are minimalist, we should "process" them by *first* developing the theory of causal taxonomies and causal similarity metrics, and *then* showing how distinctions of explanatory relevance between causes and preempted backups fall out. And it is Strevens's notion of "depth" that should serve as the touchstone for such a philosophical account of causal taxonomy and similarity.