Biol Philos
DOI 10.1007/s10539-010-9234-2

REVIEW ESSAY

Makes a Difference

Review of Michael Strevens' Depth: An Account of Scientific Explanation. Harvard University Press, Cambridge, MA, 2008

Arnon Levy

Received: 10 October 2010 / Accepted: 21 October 2010

Abstract Michael Strevens has produced an ambitious and comprehensive new account of scientific explanation. This review discusses its main themes, focusing on regularity explanation and a number of methodological concerns.

Keywords Scientific explanation · Causation · Idealization · Abstraction · Regularity explanation · Counterfactuals

In *Depth,* Michael Strevens attempts to furnish a novel and comprehensive way of thinking about scientific explanation. Remarkably, the attempt is met with success. The book is not without shortcomings, but lack of originality, breadth or systematicity are not among them. On offer is a general theory of explanation, encompassing both single-event and regularity explanation. Moreover, the core theoretical apparatus serves as the basis for a number of significant theses concerning scientific explanation, among them a view of probabilistic explanation, an account of the explanatory role of idealization and mathematical abstraction, and informative discussions of a number of issues surrounding explanation in the "special sciences", including evolutionary biology and ecology.

Strevens advocates a species of the causal approach to explanation, and one of the key respects in which his theory is novel is that it takes on very minimal commitments regarding causation. Most casual accounts of explanation are grounded, in some way or other, in a view about the metaphysics of causal relations. Cuasalists of different sorts, including Salmon (1984), Lewis (1986) and Woodward (2003) are concerned primarily with the kind of relation that holds

A. Levy (✉)
The Van Leer Jerusalem Institute, 43 Jabotinski St., 91040 Jerusalem, Israel
e-mail: arnondor@gmail.com

Published online: 01 March 2011

between two events or variables that we think of as cause and effect–such as counterfactual dependency or the conservation of a fundamental quantity. The move from causation to explanation is then seen as straightforward: All causes are taken to be, at least in principle, fit to play a part in explanation, with perhaps some low-key selection criteria operative in specific contexts. In contrast, Strevens opts for a "two-factor" theory, which recognizes an important difference between causation and causal explanation. On this kind of view explaining consists in *selecting amongst causal influences only those that are explanatorily relevant* to understanding some phenomenon of interest. Thus, the discussion is centered on specifying criteria that set apart the causes that play a part in explanation from those that don't.

One place where the two-factor approach marks a substantial shift is in the treatment of higher-level causal relations. Many views on causation accept that there are non-fundamental causes in part because of the role of non-fundamental entities and events in scientific explanation. It is then seen as a constraint on any (metaphysical) account of causation that it allow for higher-level causes. Strevens, relying on the two-tiered structure of his view, acknowledges only fundamental level causal influences and treats higher level causal explanations as, essentially, judgments about explanatory relevance—distillations of the relevant causal influences. This reflects a bias towards physically fundamental entities and events, which might not be to the liking of some of his readers, especially those in the philosophy of biology. On the other hand, Strevens does not deny the importance of non-fundamental causal claims and his account of them (on which more anon) comports well with his overall view of causal relevance. That account has independent considerations supporting it, and these might lend indirect support to Strevens' "fundamentalist" bias.

Another feature of a two-factor approach is that, because explanation is seen as one step removed from causation, the theory of explanation can take on a far more modest metaphysical character, remaining neutral on what, exactly, causal relations are. However, such a conception does raise questions regarding the subject matter of a theory of explanation; I come back to this issue below.

At the heart of Strevens' theory—which he refers to as the "kairetic" theory—is a procedure for isolating explanatorily relevant causal influences. This isn't meant as a descriptive account of how scientists construct explanations, but as something like a rational reconstruction of the criteria they employ. We start with what Strevens calls a "causal model" for a phenomenon to-be-explained: a set of statements that entails a statement of the phenomenon, while mirroring the causal process at work. (Relatively little is said about what such mirroring consist in). Once we have a causal model in hand, we engage in a process of abstraction: we "cleanse" the model of details that are not necessary for the causal entailment of the explanandum—either eliminating propositions that are unnecessary, or else replacing them with less detailed alternatives (e.g. instead of stating a precise value for some variable, we state that it must be within a certain range). Doing away with all unnecessary details results in an abstract model, consisting only in those factors without which the model does not entail the explanandum. In an obvious and quite precise way, these are factors that the model could not do without. Strevens regards them as "difference makers" for the explanandum.

Thus, the first and key requirement is that an explanatory model be as abstract as possible, i.e. that it omit all possible detail while not invalidating the relevant causal entailment. But abstractness is not the only constraint on an explanatory causal model. It must also be cohesive—if a model can be instantiated in several ways, then a cohesive model is one in which the instantiating causal processes are similar to each other, not an agglomeration of unconnected causal pathways. Cohesion is Strevens' version of the idea that a unified account is explanatorily superior to a disunified one. This notion is most familiar from the unificationist strand in the literature on explanation, primarily Kitcher's (1989) treatment. But whereas Kitcher saw unification as a formal notion, essentially the requirement to use as few representational schemas in explanation as possible, Strevens treats cohesion as an ontological desideratum. He presumes that causal processes can be ordered, at least roughly, in a similarity space, such that it makes sense to think that some are more similar to each other than others (§3.63). This presumption is less obviously true then Stevens seems to suppose, especially when the dimensions of the relevant comparisons are left unspecified. Moreover, even if one accepts the notion that causal processes can be ordered in terms of similarity at the level of fundamental physical dynamics (as Strevens proposes) it is unclear how well this similarity space corresponds to our judgments in the domain of explanation. That said, it seems that the basic idea, namely that abstractness and cohesion are central virtues in explanation, has much going for it.

The core of the Kairetic view, then, is that a good explanation is the most abstract and cohesive causal description of a phenomenon. Abstraction is understood, basically, as the omission of as much detail as possible, whereas cohesion is taken to be similarity of the represented causal processes. The most abstract cohesive model is one that, according to Strevens, contains only "difference makers" and thus is suitable for explanation. Strevens argues for the adequacy of this view primarily by demonstrating that a model that meets the specified constraints fits intuitive judgments about high-level causation (recall that, on the "two-factor" approach, high-level causal judgments are really judgments about explanatory relevance). He argues that the Kairetic procedure provides the right results in scenarios exhibiting causal preemption, over-determination and related phenomena. These kinds of scenarios, and the problems they pose for various accounts of causation and explanation, have been much discussed in the literature on causation. I won't attempt to survey or assess Strevens' success in accounting for the puzzles in this area. But I think it is worth commenting on a methodological issue, namely, that it isn't obvious that examining an account of explanation against intuitive causal judgments is the right kind of test for a theory of *scientific* explanation. No doubt situations involving preemption and over-determination are frequently explored by scientists. It is also likely that the question how well a candidate explanation handles a preemptive or redundant process may sometimes enter into judging the worth of the proposed explanation. But it is not obvious that the judgments of scientists and armchair-dwelling philosophers concur on these questions. The judgements of a geneticist working on, say, a regulatory network in which there is functional redundancy may diverge quite considerably from the views of philosophers reflecting on the type of modifications of everyday causal scenarios that fuel discussions of causation. To clarify: this is not meant as a complaint about the methodology of metaphysicians, but about its suitability to a project that seeks to elucidate scientific practice.

Strevens introduces the basic Kairetic account in the context of single-event explanation. This is standard practice. Indeed many discussions of explanation have had very little to say about the explanation of lawful generalizations and other regularities. But Strevens devotes substantial attention to regularity explanation. Generally speaking, he holds that the same principles govern both event and regularity explanation. This idea is encoded in what he labels, rather audaciously, the "fundamental theorem" of explanation: "the explanation of a causal generalization and the explanation of any instance of the generalization invoke the same causal mechanism." (p. 223).[1,2] But beyond this basic uniformity, there are two important ways in which explaining events and explaining regularities differ.

The first involves, unsurprisingly, the role of generalizations. Strevens contends that every regularity explanation relies on specifying a causal mechanism, but also on a "basing generalization"—a physically contingent attribution of a property to a set of objects. Thus, the (physiological) explanation of why all ravens are black contains an account of the mechanism that generates dark pigmentation in feathers (call that mechanism M) plus a basing generalization to the effect that all ravens (or at least the vast majority of them) are equipped with M. (The property need not be the having of the mechanism. It might be a property that results in the activation of a mechanism that exists more widely). Thus, a basing generalization says that the objects in a certain domain of interest share a certain structural or mechanistic feature (which in turn has the causal power to generate an instance of the explanadum[3]). Together, the basing generalization and the mechanism explain why objects in the domain have the property to-be-explained.

Importantly, however, not any generalization can serve as a basing generalization. Only generalizations that state a pattern of "entanglement", to use another bit of Strevenese, will do. Entanglement is a robust counterfactual connection, in more or less the following sense. If F is entangled with G then, within a certain normal range of changes and manipulations, changing the object in a way that does not affect its F-ness will not affect its G-ness. And, in addition, changes that do affect the objects F-ness will tend to affect (or altogether eliminate) its G-ness as well. Thus, to say that F is entangled with G is to say that they go together in not-too-unusual counterfactual circumstances. The notion of entanglement is thus a reach out, on Strevens' part, towards the counterfactual approach to explanation.

Strevens argues that entanglement is an explanatory relevance relation. This is viewed as an important new insight. Indeed the cover flap announces that one of the main achievements of the book is that it "augments the familiar causal approach to

[1] This is in fact the first of two *fundamental theorems*. But the second (stated on p. 260) is similar in content.

[2] The term "mechanism" is used by Strevens to refer rather loosely to causal generalizations. There is no specific connection here with the debate over mechanisms in the philosophy of biology.

[3] This last clause underlies Strevens' fundamental theorem. The explanation of an event is the explanation of the regularity, minus the basing generalization, as it were.

explanation [with] a new, non-causal explanatory relevance relation—entanglement".[4] Setting aside the question of what is involved in calling something an "explanatory relevance relation", one wonders how much of a step forward the notion of entanglement represents, without an underlying account of the role of counterfactual information and some indication of the relevant epistemology. This is an issue that Strevens ought to be sensitive to. In earlier parts of the book he criticizes counterfactual approaches to causal relevance, especially the manipulationist theory advocated by Woodward. The discussion of Woodward is rather condensed and not always easy to follow. It seems that Strevens' main complaint is that, in the manipulaitonist view, explanations of particular events, including non-fundamental ones, are reliant on judgments about the manipulability relations between types of events. Such type-level relations, in turn, do not receive any elucidation by Woodward. Now, it seems to me that this argument is not entirely fair to Woodward, whose goal is not to supply a top-to-bottom theory of casual relations, but to show how to take a casual network we have some information about, and figure out which events in it are causes of others. This kind of project places some distance between the metaphysics of causation—at any rate, the fundamental metaphysics of the matter—and questions about causal judgments. In this regard Woodward's project is not dissimilar to the one undertaken in *Depth*.[5] Be that as it may, it seems that Strevens' advocacy of entanglement is open to a criticism analogous to the one he directs at Woodward. If entanglement is a genuine explanatory relevance relation, then it would seem that Stevens must tell us how to detect patterns of entanglement—where does knowledge of type-level relations of entanglement come from? Strevens gestures in this direction when he defines entanglement as a robust counterfactual relationship. But a bare appeal to counterfactuals—especially in the present context—raises many questions and supplies few answers. Thus, if the claim made against Woodwardian manupulability relations are sound, its analogue ought to be at least as worrisome with respect to Strevensian entanglement.

Setting such epistemic worries aside, it does seem that much of the point of invoking a pattern of entanglement in an explanation is that it embodies counterfactual information, including information about potential manipulations, about the explanans. In appealing to entanglement, therefore, Strevens' account takes a rather substantial step in the direction of a manipualtionist, or more broadly a counterfactual-based account.

The second topic that Strevens addresses in the context of regularity explanation is idealization. Broadly speaking, idealization involves the deliberate misrepresentation of natural phenomena: supposing that a plane is frictionless, that a particle is a perfect sphere or that a population of organisms is infinite. The ubiquity of idealization, coupled with the sense that it often enhances understanding, poses a

[4] It is not entirely clear what is involved in calling something an explanatory relevance relation. Presumably it means that an appeal to this kind of relation in the course of providing an explanation may enhance its explanatory power. However, this might be true of various logical relations as well, depending on the context.

[5] On this score see the somewhat heated yet highly illuminating exchange contained in: Strevens (2007), Woodward (2008) and Strevens (2008).

special problem for causal views of explanation, as Strevens notes. For idealization often involves a distortion of the casual process involved, and so seems to undermine the explanatory power of a model rather than enhance it. There has been a lot written on idealization in the last several decades in the philosophy of science, especially in the context of physics. But general treatments of the role of idealization in explanation are, perhaps surprisingly, few and far between. Strevens tackles the issue head-on. He suggests that the role of idealization is to convey information about difference making factors. Idealization, he argues, contributes to explanation when the factors that are misrepresented are ones that do not make a difference to the explanandum. Taking the perspective of the Kairetic account, this means that idealization ought to be performed when one can abstract away from a factor without loss of causal entailment. But the more basic idea Strevens offers is independent of his particular brand of the causal view of explanation. The basic idea is that idealization is kosher when, and only when, what is idealized is causally inessential to the explanation. The role of idealization is then precisely to signal that the factor(s) in question are of this sort—the non-difference-making sort. Thus idealization is depicted as having a communicative role in explanation. In presenting an idealized model, a scientist communicates a kind of negative causal information.

This an attractive idea, and it presents idealization as not only compatible with causal explanation but as enhancing it. Although gestures towards this sort of conception of idealization have been made in the literature, I think Strevens is the first to spell it out and argue for it. However, it is not clear how general such a communicative account of idealization really is. Most idealized models do not meet the standard that Strevens sets them: they misrepresent factors that do make a difference. Indeed it is arguable that idealizing assumptions are often employed in a manner that requires the sacrifice of difference-making information. This, it seems, is because they advance other explanatory goals, such the desire to make some factors salient at the expense of others, or the construction of a general model that encompasses a variety of phenomena. Stevens does not address these options directly and it is not entirely clear to me whether these are compatible with his account. Perhaps he thinks that idealizing in the name of salience or generality is fine but unimportant. But at some point he seems to regard the conveying of negative causal information—information to the effect that some factor does not make a difference—as the sole role of idealization. Part of the reason, I think, has to do with the kinds of examples that Strevens takes as primary. These involve assigning "default values" to key variables: treating friction as null or population size as infinity. In these cases, or at least in some of them, it is easy to regard idealization as signaling that friction or population size makes no difference. But in other cases the matter is not so clear. How to think about a model that idealizes away intergenerational overlap (common in population biology texts)? Or one that treats selection as operating at a single genetic locus? Not only do these not look like "default values" in the sense alluded to above, they also, in many cases, appear to involve factors that do—contra Strevens' main thesis—make a difference. Thus, it would seem that the Strevens account is either limited in scope, or has the consequence that many real-world idealized models are explanatorily inferior to (potential or actual) non-idealized counterparts. It is doubtful that this last option reflects scientific practice and common standards of assessing success in explanation.

So much for regularity explanation. The next-to-last part of the book, Part IV, is devoted to probabilistic explanation.[6] It is long and involved, and I will not be able to do it any kind of justice here. Strevens distinguishes several kinds of roles that probability plays in explanation. He also discusses questions of size and direction—does the fact that a factor makes to an outcome more or less probable, and the degree to which it does so, matter to its explanatory relevance? Perhaps the main thesis argued for in this part of the book is the surprising idea that the best explanation of a deterministic phenomenon is often probabilistic. Here, as in previous parts of the book, the systemticity of Strevens's thinking about explanation is striking. The arguments that are brought for many of the ideas in this section derive quite directly from the views developed earlier regarding the importance of abstraction in explanation and the role of entanglement. One disappointing aspect of the discussion of probabilistic explanation, however, is that it proceeds with very few real-world scientific examples. The discussion makes scattered allusions to cases from various sciences, but none are discussed at any length. Strevens argues for and illustrates his main points with reference to a clean, hypothetical case of a "wheel of fortune". He states that this case is representative of many real-world cases,[7] but for a proof of this point he sends the reader to his earlier work on complex systems (Strevens, 2003).

A dearth of scientific examples characterizes other parts of the book, and represents one of its main drawbacks. Strevens introduces the Kairetic account with the aid of the story of the murder of Rasputin—the late 19th century Russian "mad monk" who was murdered in legendary and causally tortuous fashion by a cabal of tsarist noble-men. Rasputin's demise involves preemption and over-determination sub-plots. As noted earlier, the reliance on thought experimentation might disappoint readers who are looking for an account that engages closely with science. But the choice of a-prioristic methodology in this regard is odd even in Strevens' own terms. Early on, he states that the goal of the book is to provide "a description of actual scientific explanatory practice" (p. 37). Several potential sources of evidence for such an account are enumerated. There are two that Strevens appears to regard as particularly important. The first is "the sum total of the explanations regarded as scientifically adequate in their day, together with an understanding of the background against which they seemed adequate" (p. 37). "The second most important source of evidence concerning our explanatory practice" we are told, "is introspective reports on the principles used in assembling particular explanations." (p. 38). The appeal to introspective information is presumably linked to thought experimentation and hypothetical reasoning undertaken by the author and his readers, rather than to interviews with scientific practitioners or similar psychological and sociological data. In the course of the

[6] The last part of the book is brief and more tentative, and I will skip it entirely here.

[7] Specifically, it is "microconstant" and "macroperiodic", two properties of deterministic systems that, together, Strevens argues, ensure the propriety of a probabilistic explanation.

book, Strevens relies primarily on the second source of evidence, especially when introducing the core of the Kairetic account (the main exception is the discussion of idealization, in which a text-book explanation of Boyle's law is examined at some length). Now, while it is likely that there is continuity between "folk" intuition and scientific practice, it is also likely that the two diverge. The complaint here is not that a reliance on hypothetical examples has led Strevens astray. Perhaps in some places it has led him away from interesting issues. But he has kept on track, I think, in the main junctures. What is missing, rather, are arguments that will convince those whose work focuses on and whose sensibilities derive from close observation of the practice of science.

Another consequence of this methodological choice is that Strevens fails to give instruction on how to scale, weigh and combine the various desiderata he proposes. In principle, one of the attractive features of the kairetic approach is that it allows explanatory success to be a matter of degree. Strevens sees this and often speaks in terms of one explanation being better then an alternative in one respect or another. But he is rarely explicit about how to measure explanatory success, and especially, about how various aspects of explanatory value—such as cohesion and abstractness, for instance—ought to be traded off against each other in arriving at an overall assessment of the quality of an explanation. It is likely that an engagement with actual scientific examples, in which such tradeoffs are commonly performed, would have led to a more explicit discussion of these important questions.

This connects to another general concern, having to do with the two-factor approach. At several points Strevens notes that it is possible to interpret the conditions he proposes for explanatory adequacy either as criteria for explanatory relevance or, alternatively, as criteria for higher-level causal relations. As he notes, such a move can be done with respect to all major accounts of causal explanation—such as the manipulability account, or the conserved quantity account. One can treat them as proposals for the characteristics of casual relations (as they are intended) or, if one wants, regard them as criteria for selecting, amongst causal influences, those that have explanatory relevance, (the second "tier"). Strevens favors the latter approach primarily, it seems, on the grounds that it entails few if any metaphysical commitments. But while an account that focuses on explanatory practices entails few metaphysical commitments, it does carry other kinds of commitments. The main one has been noted already: an account of our practice is, effectively, a description of the principles that guide scientists in constructing and evaluating explanations. Providing an account of this sort entails empirical commitments. Questions such as "how do scientists (in general, or in some area) determine what counts as a good explanation?", "how do they rank explanations?" etc. are obviously empirical questions, at least in part. Even if one believes, as Strevens does, that answers to these questions remain essentially fixed over time and among different cultures, an account of this sort must of necessity take on substantial empirical commitments. Thus, while a two-factor approach relieves the philosopher of one kind of commitment, it generates another, a kind of empirical adequacy. Strevens often fails to show that his account is empirically adequate.

Having said that, it should be clear that the choice to work out a two-factor approach also marks one the main attractions of this book. Overall, Strevens charts a course that successfully avoids the intellectual remoteness that sometimes besets the metaphysics of causation while managing to offer a general, systematic and novel approach to explanation. This is an achievement that only a handful of contemporary philosophers of science can take pride in.

References

Kitcher P (1989) Explanatory unification and the causal structure of the world. In: Kitcher P, Salmon WC (eds) Scientific explanation. University of Minnesota Press, Minneapolis

Lewis D (1986) Causal explanation. In: Philosophical papers, vol 2. Oxford University Press, Oxford

Salmon WC (1984) Scientific explanation and the causal structure of the world. Princeton University Press, Princeton

Strevens M (2003) Bigger than Chaos: understanding complexity through probabaility. Harvard University Press, Cambridge

Strevens M (2007) Review of Woodward, making things happen. Philos Phenomenol Res 74(1):223–249

Strevens M (2008) Comments on Woodward, making things happen. Philos Phenomenol Res 75(1):171–192

Woodward J (2003) Making things happen. Oxford University Press, New York

Woodward J (2008) Response to Strevens. Philos Phenomenol Res 75(1):193–212