Ceteris Paribus Hedges: Causal Voodoo That Works

Michael Strevens

Journal of Philosophy 109, 652–675, 2012

What do the words "ceteris paribus" add to a causal hypothesis, that is, to a generalization that is intended to articulate the consequences of a causal mechanism? One answer, which looks almost too good to be true, is that a ceteris paribus hedge restricts the scope of the hypothesis to those cases where nothing undermines, interferes with, or undoes the effect of the mechanism in question, even if the hypothesis's own formulator is otherwise unable to specify fully what might constitute such undermining or interference. I will propose a semantics for causal generalizations according to which ceteris paribus hedges deliver on this promise, because the truth conditions for a causal generalization depend in part on the—perhaps unknown—nature of the mechanism whose consequences it is intended to describe. It follows that the truth conditions for causal hypotheses are typically opaque to the very scientists who formulate and test them.

1. Regularity and Hedge

High-level regularities—patterns of biological, psychological, social, or economic events—are, in our kind of universe, convoluted things. For example, a government's printing money to pay its debts tends more often than not

to cause an increase in the rate of inflation. But under a number of different circumstances, printing money may not affect inflation (if, for example, the extra currency is hoarded in mattresses rather than spent, so taken back out of circulation). Because these circumstances are rather diverse, an attempt to specify the economic regularity with any degree of precision will be a daunting undertaking, requiring presumably many clauses, subclauses, parentheses, and footnotes. Some writers have thought it plausible that an exact specification of the regularity would have to be infinitely long.[1]

The reasons for the convoluted nature of high-level regularities are many, but one is far more important than the rest: the high-level regularities that interest science tend to be the consequences of somewhat complex causal mechanisms, and such mechanisms are not even close to being sure-fire things, for three reasons. First, they require certain enabling conditions to operate, but these conditions are not always present. Second, even when the conditions are right, the operation of a mechanism is always vulnerable to interference from the outside, to some orthogonal force that derails the causal train of events. Third, even when a causal process runs to completion, other causal processes may reverse or otherwise undercut its effects. (Not all dogs are four-legged in part because some are amputees.) To capture precisely the consequences of a causal mechanism, then, you must capture all such possibilities for causal failure or reversal. An exact specification of such conditions, and so of the regularity itself, must be if not infinite then at least extremely long and complex.

This complexity—this intricacy in the twists and turns of the high-level regularities—poses a problem, it is generally agreed, for our understanding of the nature and workings of the high-level sciences. To introduce the problem, suppose that a statement of a high-level law should entail its corresponding Humean generalization, where by a law statement's "corresponding Humean

1. Paul Pietroski and Georges Rey, "When other things aren't equal: Saving ceteris paribus laws from vacuity", *British Journal for the Philosophy of Science* 46, 81–110, 1995, p. 102.

generalization" I mean a precise specification of the pattern of events to which the law in question would give rise, if the law statement were correct. The law statement *All Fs are Gs*, for example, should entail that all actual *F*s are *G*s—which is why an *F* that is not *G* is regarded as a counterexample to the putative law. (On the view that a law statement is nothing over and above a Humean generalization, the entailment principle is trivially true.)

From the entailment assumption it follows that law statements must be at least as complex as their corresponding Humean generalizations. Consequently, if all (interesting and correct) high-level Humean generalizations are extremely long, complex, and convoluted, all true high-level law statements must be semantically at least as convoluted. Two methodological problems then arise.

First, if the degree of complexity is high, it is hard to see how researchers in the high-level sciences can formulate law statements that have any chance of being correct. Because coming to know the nature of the high-level laws requires formulating, then testing and confirming, correct high-level law statements, it follows that it is practically impossible for science to discover the high-level laws.

Second, even if it were possible to formulate law statements of sufficient length and complexity to specify the high-level laws, it would be a formidable task to discover which of many similar complex law statements was correct. For any detail that might plausibly appear in the specification of the law under investigation, there will be two statements that differ only on that detail. Determining the correct law statement would seem to involve as many tests as there are details.

There are several possible responses to these worries. First, you might give up on the idea that the sciences, or the high-level sciences at least, take as one of their principal goals the discovery of laws.[2] It is not enough, note, merely to collapse the distinction between laws and other generalizations, since the

2. Ronald N Giere, *Science without Laws*, University of Chicago Press, Chicago, 1999.

problems stated above for laws look to apply equally to any class of scientifically significant generalizations. This first response, then, is quite radical: it is to abandon the scientific search for high-level regularities altogether.

A second response is to understand high-level law statements (or generalizations, or whatever you want to call them) in such a way that they do not entail their corresponding Humean generalizations. You might, for example, think that the purpose of a high-level law is to assert the existence of a tendency or a capacity.[3] Though such a capacity will give rise to a complex pattern of events, the assertion of its existence need not logically entail this pattern.

A third response is to understand high-level laws in such a way that their corresponding Humean generalizations need not capture the convolutions described above. High-level laws, you might hold, are "statistical" or "inexact";[4] their statements entail the existence of certain patterns of events, but these patterns are themselves rather coarse-grained—coarse-grained enough to accommodate, without closely following, the empirical signature of causal breakdown. Perhaps the corresponding Humean generalizations are vague, or perhaps they allow for many exceptions ("Most *F*s are *G*s, most of the time").

A different way of achieving the same end is to understand high-level laws as making claims not about the real world but about idealized systems or models in which Humean generalizations are relatively simple, and whose similarity to the real world is intended to be only approximate.[5]

It is a fourth response that is the subject of this paper: there is (so the response goes) a certain generic hedge—in this paper, I will suppose that it is the expression "ceteris paribus"—that, when added to a law statement, imbues the statement with additional content in virtue of which it entails

3. Nancy Cartwright, *Nature's Capacities and Their Measurement*, Oxford University Press, Oxford, 1989 and Peter Lipton, "All else being equal", *Philosophy* 74, 155–168, 1999.

4. Daniel M Hausman, *The Inexact and Separate Science of Economics*, Cambridge University Press, Cambridge, 1992.

5. Also a view associated with Cartwright.

its corresponding generalization without its having to specify explicitly the generalization's convolutions. For example, although *Printing money causes inflation* entails only the simple and therefore false Humean generalization *Printing money is always accompanied by a rise in inflation*, the addition of a ceteris paribus hedge changes everything: *Ceteris paribus, printing money causes inflation* entails a far more nuanced Humean generalization that takes into account all the opportunities for the causal mechanism linking money printing and inflation to be stymied by the absence of enabling conditions, interference, or reversal. I emphasize that this is a thesis about the content added by the linguistic item "ceteris paribus" to a *law statement*, not a view about the laws themselves. It is unremarkable to hold that the actual laws of nature necessitate real rather than fictional patterns of events; what is intriguing is the possibility that, with the help of a familiar Latin expression, we can frame short and simple sentences that entail actual event patterns in all their glorious and gory complexity.

Such a view implies that our hypotheses about the laws—our attempts to formulate true law statements—typically have what I will call opaque conditions of application. By a law statement's *conditions of application* I mean simply a rider on a law that restricts the range of systems to which it applies. For example, in classical genetics' law statement *If two genes lie on different chromosomes, they assort independently*, the condition of application is the "if" clause, restricting the scope of the law to genes on different chromosomes. More generally, in a statement of the form *In conditions Z, Fs are G*, the conditions of application are *Z*.

A law statement's conditions of application are partly *opaque* if they are not all known to the scientists who are testing or otherwise putting the statement to use. If the addition of a ceteris paribus hedge to a law statement amounts to a requirement that the causal mechanism operate unimpeded, it will in almost every case add opaque content to the statement, because the statement's user—a scientist who is perhaps just beginning to investigate the nature of

the mechanism in question—will normally be unable to specify explicitly a complete set of conditions sufficient for the mechanism to function, or roughly equivalently, will be unable to recognize in all cases whether or not such conditions hold.

This is a rather loose characterization of opacity. There are in fact many ways that a notion of opacity might be defined—that is, there are many related opacity-like properties. Any of these would serve to make my point in this paper, so there is no need to construct a precise definition. What matters is that these notions of opacity share a certain methodological consequence: that a hypothesis about a law is opaque implies that the scientific community is not always in a position to recognize when that hypothesis's conditions of application hold, because the community does not in the methodologically relevant sense grasp those conditions' entire content.

Let me add that opacity of this sort should not be unfamiliar: it will exist wherever hypotheses use vocabulary that is "externalist" in a certain sense, as natural kind terms are usually thought to be. For example, if a scientist formulates a hypothesis about gold without fully understanding the nature of gold, they may not be able in all circumstances to distinguish even in principle whether their hypothesis applies to a certain specimen, because their theoretical knowledge will be unable to decide the question whether the specimen is in fact a sample of gold.

Opacity raises two questions. First, it may seem too good to be true that, by adding the words "ceteris paribus" to your hypotheses about the laws, you can endow them with a sophisticated causal know-how that is otherwise beyond your reach. Second, this power of ceteris paribus hedges may seem to be not only miraculous but useless. What is the practical significance of content in a hypothesis unless the investigators know that it is there?

The aim of this paper is to address the first worry, by showing that there is a simple and natural way of understanding the semantics of law statements on which they can have the kind of opaque content advertised above.

* * *

The account I give of the power of ceteris paribus hedges is limited to their role in *causal generalizations*, by which I mean generalizations that are intended to describe certain patterns of events that are consequences of the operation of some single kind of causal mechanism, which I call the hypothesis's *target mechanism*. (It is the actual facts about this target mechanism, I will argue, that determine the significance of the hedge.)

For simplicity's sake, I focus here on causal generalizations of the form *Fs are Gs* or *F-ness causes G-ness* that are presumed to hold in virtue of a causal mechanism connecting *F*-ness to *G*-ness. My generalizations are directional, then: they attribute to the antecedent an active role in a causal process that brings about the consequent. (Contrast a generalization that associates two effects of a common cause such as *Barometer drops are followed by storms*, or a generalization that is intended to be agnostic about the direction of causation.)

A few remarks. First, as far as I can see there is no special syntactic form that marks out the causal generalizations with which I am concerned. A natural formulation uses the generic: *Printing money leads to inflation*, *Ravens are black*, and so on; causation need not be mentioned explicitly.[6] Indeed, causal generalizations are sometimes expressed without any words at all, in the form of mathematical equations presented in a context where the implicit claims of causal directionality are clear to all. Thus, although causal generalizations are a particular kind of proposition with (I will propose) a specific semantics, the linguistic and other communicative resources used in science to express such propositions are quite heterogeneous.

Second, I sometimes use the term "causal hypothesis" rather than "causal generalization", both for variation and to stress that my subject matter is scientists' attempts to represent laws rather than the laws themselves. (I will, incidentally, no longer use the term "law", which seems too grandiose in the

6. Bernhard Nickel, "Ceteris paribus laws: Generics and natural kinds", *Philosophers' Imprint* 10, 1–25, 2010.

context of many of the high-level sciences.)

Third, I assume that the causal mechanisms in question are deterministic.

2. Truth Conditions for Causal Generalizations

The canonical form of a directional causal generalization is, I will suppose, *In conditions Z, Fs are G*. (You should encounter no problems in generalizing what I have to say to other similar forms, such as *Fs are followed by Gs*, *Fs behave in manner G*, and so on.) What makes such a generalization causal is the supposition that there is a causal mechanism by which *F*-ness helps to cause the *G*-ness in question—the "target mechanism".

Let me give a few examples:

1. If a gas's density and pressure are not too high, then when its temperature is held constant, its pressure varies in inverse proportion to its volume.

2. Ravens are black.[7]

3. Paranoid schizophrenics hear voices.

4. Adult humans think about biological species in essentialist terms.

5. Hunter-gatherers share large food items with all members of their band.

What kind of truth conditions does a causal generalization without a ceteris paribus hedge have? Let me make a suggestion to get things started: the causal generalization *In conditions Z, Fs are G* means

7. In this and some of the other examples, there is reason to doubt that the antecedent property is supposed by the generalization's users to play a direct role in the putative causal mechanism in question (as explained by Michael Strevens, "The explanatory role of irreducible properties", *Noûs* 46, 754–780, 2012). Such cases require additional elements to be added to the account of ceteris paribus hedges given in this paper; these complications will be left to another time.

> There exists a causal mechanism that has as its only enabling conditions or components Z and F and that brings about G.

But that is too strong; as inspection of the examples above shows, most causal generalizations do not specify every enabling condition and component of a mechanism. Perhaps the truth conditions should be weakened like so:

> There exists a causal mechanism that has among its enabling conditions and components Z and F and that brings about G.

This semantics is too weak. It makes such generalizations almost trivially true: for just about any choice of Z, F, and G there will be some Rube Goldberg mechanism that links the three as specified.

What else might work? One possibility is to retreat from making any explicit reference to a causal mechanism in the truth conditions. The assumption that the correlation between F and G is due to a causal connection would then be exiled to the conversational context, and the generalization itself would be given truth conditions such as "When Z holds, all Fs are G".

A more interesting strategy is to move elements from the context to the truth conditions rather than vice versa. When scientists formulate a causal generalization, they typically have a certain mechanism in mind concerning which they wish to make their claim. They do not conceive of themselves as making an existential claim ("somewhere out there, there is a mechanism that . . . "), but a claim about the nature and consequences of a particular mechanism that has already attracted their interest—even if they have no knowledge of that mechanism's internal workings. Boyle's law is about the intrinsic behavior of gases (thus not about Rube Goldberg causal pathways that travel outside the gas—for example, the weasel sees that the pressure has increased and hits the switch to decrease the volume); "Ravens are black" is about ravens' natural coloration mechanism (not about the efforts of unstable ex-confirmation-theorists who go around bleaching ravens white); and so on. When a causal hypothesis is framed, then, it is supposed to make a claim about a particular, contextually determined mechanism: the target mechanism.

I propose that the truth conditions for causal generalizations make explicit reference to this target mechanism; they are as follows:

> The contextually determined target mechanism M has among its enabling conditions and components Z and F, and brings about G.

or more elegantly and roughly equivalently:

> By way of the target mechanism M, the conditions Z and the property F bring about G.

Two concerns. First, these truth conditions require some precisification. Does the hypothesis assert that F and Z always bring about G, or only that they do so under further, unspecified circumstances?

Second, the proposal raises a slew of questions about the so-called target mechanism. How are such mechanisms picked out, especially at the early stages of scientific investigation when not much is known about the subject matter? What if the scientists in question are sufficiently confused that their investigative intentions fail to pick out a mechanism, as might happen, for example, if they intend their hypotheses to describe the phlogiston-consumption mechanism, or the astrological influence mechanism? What if scientists have no particular mechanism in mind?

These issues will be addressed after I have discussed the next topic, the question of what "ceteris paribus" contributes to a causal generalization's truth conditions.

3. The Semantic Contribution of Ceteris Paribus Hedges

3.1 Approaches to Ceteris Paribus Hedges

There are, broadly speaking, three approaches to understanding the significance of ceteris paribus hedges (though the literature is large and complicated, and not every view fits neatly into the following schema). In each

case, the hedge may be understood as responding to the problem of causal breakdown—the fact that causal mechanisms may not get started, or may not run to completion, or may have their effects reversed.

On the *softening* approach, adding "ceteris paribus" to the generalization *Fs are G* loosens the connection asserted to hold between *F* and *G*. The simplest softening approach turns "all" into "most". If *Fs are G* means *All Fs are G*, for example, then *Ceteris paribus, Fs are G* means *Most Fs are G*. (Fodor suggests in addition that there be no systematicity to the exceptions.)[8] Alternatively, softening might transform *Fs are G* into *Fs have a nondeterministic tendency to be G*.[9] Either way, adding the hedge to a causal generalization implies that, because of causal breakdown, there are some *F*s that are not *G*, but does not specify or otherwise imply which *F*s these are.

On the *narrowing* approach, adding "ceteris paribus" strengthens the generalization's conditions of application, so reducing the range of systems in which the connection between *F* and *G* is asserted to hold. Thus *Ceteris paribus, Fs are G* means *In conditions Z, Fs are G*, for some specific *Z*.[10] In the case of a causal generalization, *Z* will specify (and exclude) some or all of the conditions under which the target mechanism breaks down.

On the *annotating* approach, the addition of a ceteris paribus hedge does not alter the truth conditions of a causal generalization; rather, it has a pragmatic function connected to causal breakdowns, usefully commenting on them in some respect—for example, warning the user that the corresponding Humean generalization is not exceptionlessly true, or that the generalization

8. Jerry A Fodor, "You can fool some of the people all of the time, everything else being equal: Hedged laws and psychological explanations", *Mind* 100, 19–34, 1991.

9. Harold Kincaid, "Defending laws in the social sciences", *Philosophy of the Social Sciences* 20, 56–83, 1990, Lipton, *op. cit*.

10. Hausman, *op. cit*., §8.2; Gerhard Schurz, "Ceteris paribus laws: Classification and deconstruction", *Erkenntnis* 57, 351–372, 2002; Mark Lance and Margaret Little, "Defeasibility and the normative grasp of context", *Erkenntnis* 61, 435–455, 2004; Nickel, *op. cit*.

is in some sense incomplete,[11] and perhaps promising to fix it.[12]

Only on the narrowing approach, you will observe, can a ceteris paribus hedge add to a causal generalization the content needed to entail a convoluted Humean generalization; it is a narrowing approach, then, that I will offer in what follows. (I do not thereby rule out the possibility that there is something to the softening and annotating approaches—a hedge might have more than one function.)

3.2 Truth Conditions for Hedged Generalizations

The truth conditions for *Ceteris paribus, in conditions Z, Fs are G* are, I propose, as follows:

> When conditions *O* hold, then by way of the target mechanism *M*, conditions *Z* and the property *F* bring about *G*,

where *O* is the set of conditions required for the successful operation of *M* (apart from *Z* and *F*).[13] If you compare these truth conditions to the truth conditions for an unhedged generalization, which were as follows:

> By way of the target mechanism *M*, the conditions *Z* and the property *F* bring about *G*,

you will see that the impact of the hedge is to add an additional rider, restricting the scope of the generalization to cases in which the conditions *O*

11. John Earman and John T Roberts, "Ceteris paribus, there is no problem of provisos", *Synthese* 118, 439–478, 1999.

12. In the interpretation offered by Pietroski and Rey, *op. cit*.

13. Does the proposition expressed by the hypothesis directly refer to the mechanism, as though by name, or does it pick it out indirectly or indexically, as though by way of a propositional phrase such as "the intended underlying mechanism" or "the mechanism made salient by the context of inquiry"? My articulation of the truth conditions leaves open the answer to this question; as a consequence it also leaves open the question whether the historical context in which a hypothesis was initially formulated plays a special role in fixing its target mechanism, whether the present context of inquiry is all that matters, or something intermediate. Such matters must be left for another time.

hold. This is therefore, as promised, a "narrowing" account of ceteris paribus hedges.

What are these conditions *O*, these "conditions required for the successful operation of the target mechanism"? They are, I stipulate, the minimal set of conditions necessary to guarantee that (a) the enabling conditions for the mechanism hold, (b) nothing interferes with the mechanism's operation, and (c) nothing reverses or undoes the effect of the mechanism's operation. When the operation conditions hold, then, a deterministic mechanism is guaranteed to operate successfully. (On the question whether an explicit statement of the operation conditions would be infinitely long, and whether it matters, see section 4.2.)

Typically, scientists will have only very partial knowledge of the target mechanism's operation conditions. Since these conditions form a part of the content of a hedged causal generalization, scientists literally do not comprehend much of the content of their own hypotheses.

Opacity has two notable consequences. First, the scientists testing a causal hypothesis may not know what predictions the hypothesis makes about any given *F*, because they do not know exactly what conditions must hold in order for the hypothesis to predict that a given *F* is *G*.

Second, if everything goes according to plan, a causal hypothesis will entail a Humean generalization that successfully traces the convoluted contours of a high-level regularity, even though the users of the hypothesis are themselves unable to specify those contours. Exact high-level truths can be formulated, then, from a position of relative ignorance, because in order to formulate truths with complex content, it is not necessary to be fully aware of that content. The two problems posed in section 1 are solved.

I expect you have some questions. Why think that there is a fact of the matter about "the target mechanism's conditions of operation"? Is there any evidence from science that ceteris paribus hedges work in the way I have described? Why would you want to introduce opaque content into your

hypotheses, in any case? I will answer them in reverse order: the latter two in section 3.3 and the first in section 4.

3.3 Opacity in Science

Why opacity? Or more exactly, why restrict your causal hypotheses to cases in which your target mechanism's operation conditions hold even when you cannot recognize whether or not the conditions hold? The answer lies in the functional characterization of a causal hypothesis given above: a causal hypothesis is supposed to specify the consequences of the operation of a certain causal mechanism. It follows that the hypothesis ought to make predictions only about cases in which the causal mechanism operates successfully. It is simply not in the business of specifying what happens when the mechanism does not run to completion or has its effects undone. So, if it is to perform its function efficiently, it ought not to pronounce on such cases. In short, it ought to say *When the mechanism operates, you get Fs that are G*. But if the mechanism is incompletely understood, such a claim must be opaque: the investigator does not know exactly what the operation of the mechanism consists in, and so they cannot recognize in every case whether the mechanism has operated or not.

A causal hypothesis is opaque, then, because the subject matter of the investigation is opaque. Investigators want to understand the behavior of mechanism M, so they quite appropriately formulate hypotheses that restrict themselves to describing the behavior of M. But because they do not fully understand M's nature, their hypotheses will contain a restriction that is itself incompletely understood.

What reason is there to think that causal hypotheses in science actually contain opaque conditions of application? Let me give two examples.

Ceteris paribus, ravens are black. These words are normally intended to describe the effects of the natural raven-coloration mechanism, whatever it may turn out to be. Coloration that is clearly not due to the mechanism

is therefore not considered relevant to the truth of the hypothesis: a raven bleached white is no refutation of the hypothesis, because the bleached raven's color is not due to the natural mechanism.

Imagine a group of scientists testing the raven hypothesis who come across a population of gray ravens. As far as they can tell, the ravens are apart from their color quite ordinary. They conclude that raven coloration is variable; the hypothesis of raven blackness is false.

Then suppose, decades later, that they discover that a previously unknown industrial pollutant in the gray ravens' habitat—call it ABC—blocks a metabolic pathway and so prevents the development of fully black plumage. The blockage of the pathway, further, is clearly unnatural: ravens have not previously been exposed to ABC or any other such pathway blocker in their evolutionary history; the blockage causes other developmental deficits that were beyond the scientists' capability to detect or describe when the gray ravens were first observed; perhaps grayness itself is disadvantageous in some subtle way.

Upon uncovering these new facts, will the researchers continue to regard the gray ravens as having refuted the raven blackness hypothesis many years before? No; rather, they will regard themselves as having discovered that the gray ravens were all along irrelevant to the blackness hypothesis, because the blackness hypothesis was intended to describe the consequences of the natural coloration mechanism, and the grayness of the ravens was no more a product of that mechanism than the whiteness of bleached ravens. Thus, they will regard their hypothesis as having had, at the time of the discovery of the gray ravens, an implicit restriction that excused it from predicting that these ravens would be black. In other words, they will regard their hypothesis as having had an implicit rider saying, among other things, *Provided that there is no significant amount of ABC in the environment* . . . . This is a rider that they were incapable of spelling out at the time; it therefore gave their hypothesis opaque content.

But they ought not to be surprised or alarmed at this: they intended at the

time that their hypothesis apply only to the products of a particular causal mechanism—the natural coloration mechanism—and they knew at the time that their knowledge of the mechanism was incomplete. Thus they were all along aware that they had given their hypothesis a rider, the methodological significance of which they only partially grasped.

As another example, consider the hypothesis that the absolute magnitude M of Cepheid variable stars is related to their period P by the formula $M = -2.81 \log P - 1.43$. (To put this into the canonical "Fs are G" form, think of "being a Cepheid" as the antecedent property F, and satisfying the mathematical relation as being the consequent property G.)

For many years a numerical relationship of this sort was known, but the actual mechanism responsible for Cepheids' variation was unknown. Suppose that during this interlude, a variable star is found that fits the profile for Cepheids (in its spectral type, the qualitative aspects of its variation, and so on) but that does not fit the formula. Is this star a counterexample to the hypothesis? At the time of its discovery, it may certainly seem so. But now imagine that it is decades later, and it has become clear that the variability of the star in question is caused by a mechanism different in certain ways from the mechanism underlying the variability of the classical Cepheids. Something like this has happened several times in the history of the study of variable stars; in each case, once the causal facts became known, the quantitatively anomalous star was treated not as a counterinstance to the period/magnitude hypothesis but as a new kind of variable star lying outside the scope of the hypothesis.[14]

There are various ways to explain this methodological trend. You need not appeal to ceteris paribus hedges; you might rather propose that the term

14. The facts are not quite as simple as in my hypothetical example, but some close approximations would include among others: the separation of the type II Cepheids from the classical Cepheids, the separation of the δ Scuti and SX Phoenicis variables from the RR Lyrae variables, and perhaps to some extent the separation of the RR Lyrae stars from the Cepheids.

"Cepheid" was all along intended to apply only to stars possessing the same causal mechanism underlying the variability of a certain "baptismal group" of stars with respect to which the term was introduced. On any such explanation, however, the original hypothesis's conditions of application are partially opaque, and the opacity is contributed by the unknown nature of an actual causal mechanism. This is what matters: I contend that it is to these opaque conditions of application that ceteris paribus hedges point.

3.4 Quantifying

Before I consider the question of how target mechanisms are determined, let me discuss two subsidiary topics: first, in this section, the matter of the proper quantifiers to use in causal hypotheses, and second, in the next section, the question whether ceteris paribus hedges render causal hypotheses trivially true.

Although I have favored a generic form for causal hypotheses—*Fs are G*, *Ravens are black*, *Schizophrenics hear voices*—it is often natural to use a quantifier in the formulation: *All ravens are black*, *Many schizophrenics hear voices*, and so on. Is there a rule determining which quantifier to use, and perhaps when to use a quantifier at all?

Let me give an answer for one particular kind of case. Suppose that the target mechanism for a causal hypothesis *Fs are G* is deterministic, by which I mean that the mechanism ensures the *G*-ness of any *F* provided that its operation conditions hold. Is it appropriate, in such circumstances, to say *All Fs are G*?

Not necessarily. Consider the causal hypothesis *Sperm fertilize eggs*, true in virtue of the natural fertilization mechanism in the relevant creatures. Arguably, the mechanism is deterministic, or close enough: when the conditions of operation for the fertilization mechanism hold, a sperm is almost certain to succeed in fertilization. Yet it would be wrong to say that *All sperm fertilize eggs*, not because the natural mechanism is indeterministic, but because

certain conditions of operation for the natural mechanism seldom hold.

I propose that the correct quantifier is to be determined as follows. Divide the mechanism's operation conditions into two sets, those made explicit in the hypothesis and those left inexplicit (which will include the opaque conditions but also, typically, some known conditions considered too obvious or tedious to spell out). Then use a quantifier that expresses the frequency with which, when the explicit conditions are satisfied and an *F* is present, the inexplicit conditions are also satisfied.[15] Such a quantifier will capture the *apparent* strength of the hypothesis, the percentage of *F*s that are *G* when the explicit conditions of application are satisfied. This explains what is wrong with *All sperm fertilize eggs*: it misquantifies the frequency with which the inexplicit conditions hold. What you ought to say, if you feel the urge to quantify, is that *some* sperm fertilize eggs.

This is barely a start to the topic of quantifiers. In the indeterministic case, should the quantifier quantify apparent strength or physical probability, or a mixture of the two? How is the question of quantification connected to the semantics of generics?[16] These matters I leave to another time.

3.5 Are Hedged Generalizations Trivial?

Could it be that, by contributing conditions to a causal generalization that guarantee the successful operation of the target mechanism, a ceteris paribus hedge trivializes the generalization? (The grandfather of all such worries is the concern that *Ceteris paribus, Fs are G* means *Fs are G, except when they are not*, in which case a hedge transforms its generalization into an empirically vapid analyticity.)

15. If there are parts of the mechanism—as opposed to operation conditions—whose presence is not entailed by the presence of *F*, then the quantifier should also take into account the conditional frequency with which these parts are present.

16. Gregory N Carlson and Francis Jeffry Pelletier (eds.), *The Generic Book*, University of Chicago Press, Chicago, 1995.

On my view, a hedged causal generalization is in no way trivial. Hedging *Ravens are black* may protect it from some potential counterexamples—from the previous section's gray ravens, for example—but not from all. What the hypothesis says is that the natural coloration mechanism for ravens makes them black. Such a claim could well have turned out to be false. We might have discovered that all ravens are naturally gray (the black ones we saw first were suffering from a rare ailment), or that ravens come in a variety of colors. The effect of a ceteris paribus hedge is to focus a causal generalization on the consequences of a particular causal mechanism. It makes an empirically substantive claim about those consequences—perhaps, as in the case of the Cepheids, a quantitative claim. If that claim is false, the hypothesis is false. In short, while a hedge safeguards a causal hypothesis against refutation by states of affairs not caused by the mechanism in question, states of affairs that *are* caused by the mechanism may, and in many cases will, disconfirm a hypothesis.

Here is a somewhat more subtle worry. In the post-war period, many economists came to believe that there was a robust relationship between inflation and unemployment: the higher the inflation rate, the lower the unemployment rate. Call this *Phillips' hypothesis*.[17] Confidence in Phillips' hypothesis may explain national policies in the late 1960s and early 1970s of tolerating high inflation in order to boost employment (though there were other forces at work as well). The consequence was stagflation: contrary to Phillips' hypothesis, both inflation and unemployment increased. Many economists would say that the hypothesis was thereby refuted.

There is, however, a causal mechanism that has precisely the consequences stated by Phillips' hypothesis. That mechanism has as one of its enabling conditions that inflationary expectations should remain constant (a condition that is unlikely to hold in the real world if the government is manipulating

17. A W Phillips, "The relationship between unemployment and the rate of change of money wage rates in the United Kingdom 1861–1957", *Economica* 25, 283–299, 1958.

the inflation rate). But then if Phillips' hypothesis is allowed the benefit of a ceteris paribus hedge, you might think it will count as true in virtue of this mechanism. A ceteris paribus hedge makes it too easy, apparently, for a hypothesis to qualify as correct.

If Phillips' hypothesis were intended to target a causal mechanism that required constant inflationary expectations (whether the mechanism were known to have this enabling condition or not), then it would indeed be true. In reality, however, it was not so intended: it was intended to apply to the real causal mechanisms driving the post-war economy in the West, mechanisms that allowed all too easily for a change in inflationary expectations. What it says of these mechanisms is false. Thus, the hypothesis is false.

4. Picking Out Mechanisms

The power of ceteris paribus hedges to capture well-defined, finely determined sets of conditions for causal breakdown hinges on scientists' ability to pick out well-defined, finely determined mechanisms as the objects of their inquiry, and thus as the subjects of their causal generalizations. I do not claim that scientists always succeed in picking out a determinate target mechanism for their causal hypotheses—more on this below—but I do claim that they often succeed, even under conditions of considerable ignorance. How is this possible?

4.1 Examples of Mechanism Determination

Let me give some examples, starting with the Cepheid variable stars. When the generalization relating the luminosity and period of the Cepheids was first formulated by Henrietta Leavitt, nothing was known of the mechanism responsible for their variability (as noted in section 3.3). Leavitt observed that a number of variable stars in the Small Magellanic Cloud showed a qualitatively

very similar pattern of variation, "diminishing slowly in brightness, remaining near minimum for the greater part of the time, and increasing very rapidly to a brief maximum",[18] while also fitting the striking luminosity/period relation. She conjectured that variable stars in our own galaxy showing the same variation pattern owed their variability to the same mechanism, and so would also fit the luminosity/period relation. (She was at first unable to test this hypothesis, because she did not have a sufficiently reliable way of determining the distance and therefore the absolute magnitude of the closer stars. She did not know the distance to the Magellanic Cloud stars, either, but she did know that they were all roughly the same distance.)

Leavitt's luminosity/period hypothesis is, I think it is clear, supposed to characterize the consequences of a certain unknown mechanism for variability. The identity of that mechanism is determined by something like the following intention in Leavitt's and her readers' minds: the hypothesis should describe the consequences of *whatever mechanism causes the variability of the Magellanic stars in Leavitt's study*. The scope of Leavitt's hypothesis is therefore all variable stars having the same mechanism for variation as the Magellanic stars.

Such an intention is quite capable, I hope you will agree, of giving the hypothesis a determinate target mechanism whose inner workings are entirely unknown to the hypothesizer. To realize this capacity for mechanism fixing, however, three conditions must hold. First, there must be a well-defined "baptismal group" of exemplars. Second, there must be an observer-independent "same mechanism as" relation, that is, a criterion for individuating mechanisms that is capable of determining facts of the matter about which stars do and do not share a certain mechanism for variation without additional input from the formulator (or other user) of the hypothesis, who might not have a clue about the causes of stellar variability. Third, a single mechanism must in

18. Edward C Pickering, "Periods of 25 variable stars in the Small Magellanic Cloud", *Harvard College Observatory Circular* 173, 1–3, 1912, p. 1. (Article written by Henrietta Leavitt.)

fact cause the behavior of all or almost all of the members of the baptismal group—it must not be the case that there are several different mechanisms, none statistically dominant. It is of course the second of these that raises the interesting philosophical questions, questions to be tackled in section 4.2.

As a second example, consider *Ceteris paribus, ravens are black*. The hypothesis is, as I have observed, supposed to describe the consequences of the natural mechanism for raven coloration—thus, ravens painted white, bleached, or otherwise artificially colored do not qualify as counterinstances. It appears that we have a readymade phrase—"natural mechanism"—that is capable of picking out a certain mechanism as the subject matter of generalizations like the raven hypothesis, without our having much idea how that mechanism works.

How does the term "natural" function referentially? Perhaps it picks out mechanisms that have been produced by natural selection, so that a natural coloration mechanism is one that has been selected for producing a certain coloration. The term may work in this way even when used by a biological naif, just as the term "gold" picks out the element with atomic number 79 when used by a chemical naif. I will not, however, speculate further on the semantics of "natural".

We can also pick out and formulate hypotheses about the consequences of unnatural or pathological mechanisms. Consider a causal generalization from above: *Ceteris paribus, paranoid schizophrenics hear voices*. It is intended to describe the result of a causal process that is in some sense typical of paranoid schizophrenia, but which is as far as we know not natural. The mechanism is also not yet at all well understood. How do we pick it out? We identify a class of individuals sharing certain symptoms, the paranoid schizophrenics, without having any understanding of the causes of schizophrenia, and we specify that our generalization is supposed to capture the consequences of the mechanism that is responsible for the symptoms in most individuals in the group. The case is similar to the Cepheids, then, with two exceptions. First,

rather than a small "baptismal class" we begin with a large class that we hope contains all or almost all individuals in which the mechanism in question is at work. Second, we are consequently perhaps rather less confident about the possibility that all individuals in the class experience their symptoms for exactly the same reason; we therefore intend our generalization to describe the mechanism that is responsible for the symptoms in the majority of cases, but we do not require that it be a mechanism that causes the symptoms in almost all cases.

The schizophrenia generalization raises the possibility that the explicit claims of a causal hypothesis may be used to pick out its target mechanism: the mechanism that psychiatrists have in mind as the target of *Paranoid schizophrenics hear voices* might be determined by their pointing to a certain class of patients, saying "whatever mechanism causes those people to hear voices". Does this not revive the worry that opaque content makes causal generalizations trivial? In the case at hand, does not the existence of the target mechanism—given the way that it is picked out—guarantee the truth of the generalization? It does. But it does not follow that the generalization is trivially true. Indeed, it fails to be true under just the circumstances you would want, namely, if the mechanism responsible for paranoid schizophrenia does not cause its sufferers to hear voices (or if there is no such mechanism). How can this be? The generalization's failure to be true will not follow from its saying something false about its target mechanism, but from its having no target mechanism, since no mechanism satisfies the description "whatever mechanism causes those people [schizophrenics] to hear voices". This sort of defect, which is distinct from falsehood, is discussed in section 4.3; also discussed there is the situation in which the "baptismal group" contains more than one mechanism having the designated effect. A causal generalization's picking out its target mechanism in a self-regarding way does not, then, make it significantly easier for the generalization to be true; consequently, such pickings out may safely be allowed.

So far, I have discussed examples in which the ignorance of a hypothesis's original formulators is profound: Leavitt, the original observers of raven color, and twentieth-century psychiatrists had very little idea as to the structure of the mechanisms that served as their objects of inquiry. But in many cases, investigators do have some particular workings in mind. This is frequently true in economics: when economists propose a hypothesis such as *Printing money leads to inflation*, they are able to describe to some extent, if not completely, how the mechanism works; such descriptions of course play an important role in picking out the intended targets of inquiry.

4.2 Individuating Mechanisms

If you have an observer-independent individuation criterion for mechanisms—if you have a "same mechanism as" relation that can sort mechanisms based on the actual causal facts, whether or not they are known to you—then it is relatively easy to point to a particular mechanism as the subject of your investigation, even when you know very little about the workings of that mechanism. As the cases of Cepheids and schizophrenia show, you need only to find a group of exemplars, and specify that your hypothesis concerns the mechanism at work in most or all of those exemplars. Where, then, does the mechanism individuation criterion come from?

Let me begin by describing what the criterion must do. On the one hand, it must make distinctions: it must determine that the mechanisms causing luminosity variation in classical Cepheids, RR Lyrae variables, and Wolf-Rayet stars are different, so that when we say "Same mechanism as that", pointing to δ Cephei, our words pick out other Cepheids but not RR Lyrae or Wolf-Rayet types. On the other hand, it must not make too many distinctions. Every Cepheid is different—in size, in the precise details of its composition, and so on—but we do not want these differences to count for the purposes of mechanism individuation, or else when we say "Same mechanism as that" pointing to δ Cephei, we will pick out only that single star.

To specify an observer-independent criterion for mechanism individuation, then, is to specify a criterion for determining which causal facts matter and which do not in deciding the question of "sameness of mechanism". It is, in other words, to find an appropriate observer-independent standard for causal coarse-graining.

I propose that two phenomena are brought about by the same causal mechanism just in case they have the same causal explanation. The causal facts that matter for the purposes of mechanism individuation are, in other words, the explanatorily relevant facts. Differences among the Cepheids do not make for differences in mechanism because they are not differences with respect to the kinds of factors that would be cited in a correct explanation of Cepheids' pattern of variation. The differences between the Cepheids and the Wolf-Rayet stars, by contrast, include factors explanatorily relevant to their variability.

In order to find a mechanism individuation criterion, then, simply look to the literature on causal explanation, with the following desiderata in mind.

First, you need an account of explanation on which an explanation takes the form of a causal model, by which I mean a generic description of a causal process, thus of a mechanism type. Almost every proponent of the causal approach to explanation would, I think, claim that their account can be configured to satisfy this demand.

Second, you need an account that will result in an appropriate coarse-graining, thus an account on which not every causal detail is explanatorily relevant to a system's conforming to a causal generalization. The right account will, for example, give exactly the same explanation for any Cepheid's conforming to Leavitt's hypothesis, despite the small causal differences between different Cepheids. This desideratum rules out Salmon's account of explanation, on which all the causal details are explanatory.[19]

19. Wesley C. Salmon, *Scientific Explanation and the Causal Structure of the World*, Princeton University Press, Princeton, NJ, 1984.

Third, you need an account on which the facts about explanatory relevance are sufficiently observer independent to carve out the mechanisms without the close supervision of the scientific community. This likely rules out van Fraassen's story,[20] on which the facts about explanatory relevance are determined in part by a relation specified individually for each explanatory inquiry. (The mechanism-determining standard for relevance need not be entirely objective, however: a general criterion for relevance determined by nothing over and above the values of the local scientific community could do the job, provided that the criterion is sufficient to determine relevance in any particular case without further community input.)

What is left? Many accounts of scientific explanation in the literature satisfy the desiderata—to name a few, those of Lewis, Woodward, Strevens, and even, though it is not usually considered a causal theory, Kitcher's unification account.[21] On each of these views, many details of the causal history of a given Cepheid, such as the precise trajectories of molecules, are irrelevant to the explanation of its variation. Thus you have a basis for a coarse-grained scheme of mechanism individuation.

More specifically, apply any of these accounts of explanation to the task of explaining some particular Cepheid's varying in accordance with Leavitt's hypothesis. You will get a causal model that specifies just those causal factors that were explanatorily relevant to the star's fitting the hypothesis. I suggest that you will get precisely the same causal model for every Cepheid. This model may be regarded, then, as the schema for the operation of the variation mechanism in Cepheids, and so as an individuation criterion for that

20. Bas C. van Fraassen, *The Scientific Image*, Oxford University Press, Oxford, 1980.

21. David Lewis, "Causal explanation", in *Philosophical Papers*, volume 2, pp. 214–240, Oxford University Press, Oxford, 1986, James Woodward, *Making Things Happen: A Theory of Causal Explanation*, Oxford University Press, Oxford, 2003, Michael Strevens, *Depth: An Account of Scientific Explanation*, Harvard University Press, Cambridge, MA, 2008, and Philip Kitcher, "Explanatory unification and the causal structure of the world", in P. Kitcher and W. C. Salmon (eds.), *Scientific Explanation*, volume 13 of *Minnesota Studies in the Philosophy of Science*, pp. 410–505, University of Minnesota Press, Minneapolis, 1989.

mechanism.

The causal-explanatory model describes both the intrinsic properties of the relevant mechanism and its conditions of operation, that is, its enabling conditions, noninterference conditions, and nonreversal conditions. On most accounts of explanation, the model will make no distinction between an intrinsic feature and an operation condition, since most accounts of explanation do not consider the distinction to be explanatorily significant. Does this matter? It does not, because on my view, causal hypotheses treat intrinsic features and operation conditions in exactly the same way: *When conditions O hold, by way of the target mechanism M, F brings about G* says the same thing about the world as *Together, O, M, and F bring about G*. Thus, if the hypothesis is to make a prediction about some particular *F*, both the features specified by *O* and those required for the instantiation of *M* must be present. I have no need, then, for a criterion distinguishing the intrinsic features of a mechanism from its operation conditions; there need not even be a fact of the matter about the distinction. All that is required is a criterion determining the union of operation conditions and intrinsic features; the explanatory criterion satisfies this requirement nicely.

Some further remarks about the explanatory conception of mechanism. First, a mechanism in my sense need have neither spatiotemporal nor organic integrity. The explanation of, say, a stock market crash may span many disparate and otherwise unrelated events that come together to cause the crash; these converging causal chains therefore constitute the mechanism responsible for the crash. Mechanisms need not come in boxes.

Second, a remark on the need for a mechanism's operation conditions to rule out potentially interfering and reversing factors that may be heterogeneous and infinite in number. Were this requirement to make the natural expression of the operation conditions infinitely long, I do not think it would be a problem, since an explanation and therefore an explanatory model can

be infinitely large.[22] But in fact, endless numbers of interferers can be ruled out by finite explanatory models quite compactly, by specifying not what might go wrong, but what must go right. An explanatory model will, for example, derive its explanandum from an assumption that such and such a force was negligible, without attempting to list all the possible sources of a non-negligible force.[23] I expect most or all target mechanisms to be finitely specifiable in natural language, then, for the same reason that most or all scientific explanations are finitely specifiable in natural language.

Third, although the relevant physical laws are typically included in a causal model, I am inclined not to count them as either operation conditions or intrinsic features of the mechanism. (This prevents true causal hypotheses from becoming metaphysical or logical necessities, though not all philosophers would consider this to be undesirable.) Take the causal-explanatory model, then; remove the laws, and what you have left is the conjunction of operation conditions and target mechanism, concerning the causal consequences of which a causal hypothesis makes its claim.

4.3 Failures of Mechanism Determination

When everything goes well, the intentions of the scientists who use a causal hypothesis succeed in picking out a target causal mechanism, the effects of which the hypothesis is supposed to describe. But what if things go wrong?

One possible, and perhaps not uncommon, problem is for the scientist to pick out a group of related mechanisms rather than a single mechanism. Suppose, for example, that researchers frame the hypothesis *Diabetes causes hyperglycemia*, intending it to pick out a consequence of a pathological causal mechanism operating in a group of patients that includes both type 1 and type 2 diabetes sufferers. Unknown to the researchers, there are two different

22. Peter Railton, "Probability, explanation, and information", *Synthese* 48, 233–256, 1981.
23. John Earman, John T Roberts, and Sheldon Smith, "Ceteris paribus lost", *Erkenntnis* 57, 281–301, 2002.

mechanisms operating in their sample: in type 1 diabetes, the body has lost its ability to produce insulin, while in type 2 diabetes, insulin is (usually) produced normally, but cells have acquired "insulin resistance", that is, an inability to use insulin effectively.

The hyperglycemia hypothesis, then, has no determinate target mechanism. Suppose that the hypothesis is hedged; what then? A ceteris paribus hedge is supposed to restrict the scope of a hypothesis to cases in which the target mechanism's conditions of operation hold, but the two diabetes mechanisms have somewhat different conditions of operation. There are various ways to patch things up. You might, for example:

1. Understand a ceteris paribus hedge as limiting the scope of the hypothesis to systems in which the intersection of the two sets of conditions of operation hold.

2. Understand the significance of the ceteris paribus hedge as varying with its application. When the hypothesis is tested against the medical history of a type 1 diabetes patient, the hedge picks out the operation conditions for the type 1 mechanism; likewise for a test against a type 2 patient.

3. Understand the ceteris paribus hedge as adding indeterminate conditions of application to a hypothesis, or perhaps as adding the intersection of the two sets of operation conditions along with some further indeterminate conditions.

I will not advocate any one of these proposals; rather, I suggest that they all do an adequate job of capturing the methodological implications of the situation: while the foundations of the investigation are not entirely secure, and a hedged hypothesis will consequently not have quite the kind of content that its users suppose it to have (i.e., a restriction to the operation conditions for a single kind of mechanism), the hypothesis can nevertheless play a useful role in the ongoing investigation. You might compare the case to one in which

a hypothesis contains a theoretical term with indeterminate reference, such as "jade" in the early days of geochemistry (so the story goes) or "mass" in pre-relativistic physics.[24] Provided that the scope of their indeterminacy is not too wide, such terms are, far from being scientific dead weight, valuable organizational tools in the period before their limitations are revealed by the advance of scientific knowledge.

A more serious deficiency in a causal hypothesis is its having no target causal mechanism at all. Consider, for example, the case of "hysteria" in women, a common diagnosis in Victorian Europe that was thought to account for a range of symptoms, such as faintness, insomnia, and irritability, but which is no longer considered to have any medical basis. A scientist might frame the hypothesis *Hysteria causes insomnia* intending to capture a causal mechanism common to most hysteria patients, but because the symptoms of hysteria patients (where they were real at all) were due to a wide array of causal mechanisms having little in common, such an intention seems likely to fail to pick out any target mechanism.

On the semantics for causal generalizations given above, a hedged generalization has truth conditions of the form: *When conditions O hold . . . the property F brings about G*, where *O* are the target mechanism's conditions for operation. If the generalization has no target mechanism, it therefore has a gaping hole in the middle of its truth conditions. It is semantically defective. (Generalizations with indeterminate mechanisms have a similar sort of hole, but it is not so gaping and can be papered over.)

What to do? Nothing, I suggest. The sort of causal inquiry that gives rise to hypotheses with no target mechanism is seriously flawed by anyone's lights; according to my semantics, the hypotheses themselves have a flaw that reflects the fundamental problem, that scientists are investigating the consequences and structure of a putative mechanism that does not exist.

24. Hartry Field, "Theory change and the indeterminacy of reference", *Journal of Philosophy* 70, 462–481, 1973.

It is an interesting methodological question, of course, how scientists deal with such cases—how they eventually back out of such investigative dead-ends. On my view, they must do so by recognizing that many of their hypotheses are not false but semantically ill formed. This is a familiar predicament; the same must be said for any branch of scientific theory in which some theoretical terms are semantically empty, such as (on most accounts) phlogiston theory, biorhythmics, and indeed the study of "hysteria".

* * *

A final question on a different topic. Does it matter whether or not a causal hypothesis is explicitly hedged? If opaque conditions of application are apt for scientific inquiry into causal mechanisms, why not hold that all causal hypotheses have such conditions, or to put it another way, that all causal hypotheses are implicitly hedged with a ceteris paribus clause? Let me conclude by endorsing this view: the sole semantics for causal generalizations is the semantics I have given for hedged hypotheses. Opacity and all the expressive possibilities it enables permeate causal inquiry through and through.

Acknowledgments

My thanks to audiences at the British Society for the Philosophy of Science 2011 meeting, the C.U.N.Y. Graduate Center, the École Normale Supérieure, Florida State University, New York University, the University of Missouri at Columbia, the University of North Carolina at Chapel Hill, and the University of Pennsylvania. Also to Marco Nathan, Bernhard Nickel, Alexander Reutlinger, John Roberts, Stephen Schiffer, and Gerhard Schurz.