

The Explanatory Role of Irreducible Properties

Michael Strevens

Draft of April 2009

ABSTRACT

I aim to reconcile two apparently conflicting theses:

1. Everything that can be explained, can be explained in purely physical terms, that is, using the machinery of fundamental physics, and
2. Some properties that play an explanatory role in the higher-level sciences are irreducible in the strong sense that they are physically undefinable: their nature cannot be described using the vocabulary of physics.

I investigate the contribution that physically undefinable properties typically make to explanations in the high-level sciences, and I show that when they are explanatorily relevant, it is in virtue of their extension (or something close) alone. They are irreducible because physics cannot capture their nature; this is no obstacle, however, to physics' more or less capturing their extension, which is all that it need do to duplicate their explanatory power.

In the course of the argument, I sketch the outlines of an account of the explanation of physically contingent regularities, such as the regularities found in most branches of biological inquiry, at the center of which is an account of the nature of contingent, empirical "bridge principles".

Science and philosophy are fighting a battle over reduction, or so it seems. On the one hand, for any “high-level” phenomenon—chemical, biological, psychological, economic—science claims to be able to provide, in the long term if not quite yet, a lower-level explanation, and ultimately a physical-level explanation. The enormous progress that has been made towards this goal can hardly be ignored. On the other hand, philosophers have recently claimed with increasing confidence that many explanatory properties cited by higher-level sciences—being water, being a species, being a belief, being currency—are irreducible.

The aim of this paper is to show that both sides may be correct. I will characterize very strong versions of both the scientific and the philosophical claims—strong versions of explanatory physicalism and explanatory irreducibility—and I will argue that there exists an explanatory relevance relation, a conception of the explanatory role played by irreducible properties, that allows the two to coexist.

I will provide no arguments for either explanatory physicalism or explanatory irreducibility; I will rather simply suppose that both doctrines are correct and attempt a reconciliation. Further, in contrast with much of the recent literature on reductionism, this paper will not be especially concerned with functionally defined or multiply realizable higher-level properties; the explanatory role I find for irreducible properties can be played by functional and non-functional properties, by multiply and singly realizable properties, alike.

1. Explanatory Physicalism

Everything that can be explained, can be explained physically; that is the doctrine of explanatory physicalism. More fancifully, if there were a race that only spoke and thought in the language of fundamental physics, and so could not conceive of non-physical properties, they could understand the world as

well as we do. Their route to that understanding might be rather different from our own, given that we use irreducible properties to do our explaining and they do not, but the end point would be the same.

Observe that these physics-bound super-scientists are not compelled to explain everything at the lowest possible level, that is, in terms of the trajectories of fundamental particles or the values of fields at space-time points. Their explanations can abstract away from much of this physical detail, provided that the vocabulary they use to achieve the abstraction is built solely from the resources supplied by fundamental physics. They are quite capable, for example, of characterizing a system of rigid objects entirely in terms of their centers of mass, thus abstracting away from the shape and size of the objects, because the notion of a center of mass can be defined in fundamental physical terms.

Thus, explanatory physicalism is perfectly compatible with the view that the explanation of certain phenomena is best conducted at a rather abstract level, omitting those details of physical implementation that make no difference to the phenomena's obtaining—a view I have advocated myself (Strevens 2004, 2009), following Putnam (1975), Garfinkel (1981), and others. The question whether irreducible properties play an explanatory role in the sciences is quite separate, and more difficult, than the question whether high-level or abstract properties play such a role.

I do not think that there is a decisive a priori argument in favor of explanatory physicalism. The principal motivation for the view is a posteriori: throughout the history of science, whenever we have come to understand a phenomenon well, we have seen that it is to be understood physically—we have seen, if you like, that it is the physical properties of the systems in question that are doing all the work. To make the argument in any particular instance is hardly trivial. I have tried myself, attempting to show how various elements of the behavior of complex systems can be understood in terms of the behavior of their parts (Strevens 2003, 2005). But science is not yet

complete, and so the empirical case for explanatory physicalism, though extraordinarily powerful, is as yet provisional. In any case, as noted above, I will not argue for physicalism here.

2. Explanatory Irreducibility

Many properties that play an explanatory role in the high-level sciences are irreducible; that is the doctrine of explanatory irreducibility.

“Irreducible” in what sense? I want a notion of irreducibility that is very strong and that provokes as direct as possible a clash with explanatory physicalism. I therefore define a property to be irreducible just in case it cannot be defined in physical terms—just in case there is no way of using physical vocabulary, in conjunction with the usual logical, mathematical, and other syntactical apparatus, to capture the nature of the property.

There are a number of other philosophically interesting properties that have gone by the name *irreducibility*. Pereboom and Kornblith (1991), for example, say that a property is irreducible if it is neither identical to a fundamental physical property nor has causal powers that are identical to fundamental physical causal powers. They would classify *having such-and-such center of mass* as physically irreducible, then, despite the fact that an object’s center of mass is easily defined (indeed, only definable) in physical terms. Other writers are not quite so liberal in their ascriptions of irreducibility, but regard any multiply realizable property as irreducible. Many such properties can be defined physically; they are nevertheless counted as irreducible because they are in some sense “physically disjunctive” (Fodor 1997). I do not have any objections to these notions; my own goal of reconciling explanatory irreducibility and explanatory physicalism is made most interesting, however, by choosing as strong a definition of irreducibility as possible.

Are there any physically undefinable properties that do explanatory work? That is a matter of controversy, but plausible candidates include the properties

of being a gene, a biological species, a belief or desire, a representation, an item of currency (such as a twenty dollar bill), even a planet. Where there is irreducibility of the general type, many philosophers suppose, there is also irreducibility of more specific types; thus, if the property of being a belief is irreducible, then the property of being a belief that it is raining is irreducible, and if the property of being a species is irreducible, then the property of being a member of the raven species *Corvus corax* is irreducible. For the purposes of this paper, let me assume that these properties, or some subset of these properties—it does not matter which—are indeed both explanatory and physically undefinable. This assumption will not be defended.

The claim that these various types, both general and specific, are irreducible, is typically conjoined with the claim that the tokens of the types are quite physical. For example, the property of being a raven may be irreducible, but particular ravens are physical objects; likewise, the property of being a US twenty dollar bill may be irreducible, but particular bills are entirely physical. Irreducible high-level properties exist, then, in a kind of ontological shadowland; they are unphysical, yet they can manifest themselves only by assuming physical form. More prosaically, In other words, concerning these properties the philosophical consensus rejects type physicalism while accepting token physicalism. I will follow this lead.

If any reasonable number of explanatory high-level properties are physically undefinable, then explanatory physicalism is in danger: it states that physics is able to capture the essence of any high-level explanation, but if the properties cited in such explanations are physically undefinable, how could that be? Physics would have to capture the explanatory role of irreducible properties without being able to say what properties they are. An impossibility? Not at all: a property and its explanatory role, though closely related, are not the same thing, from which it follows that even if a property is itself physically ineffable, its explanatory role might be within the physicalist's grasp. This paper aims to show that irreducible properties play just such a

role in high-level explanation.

I will not, however, examine every imaginable explanatory function that irreducible high-level properties might have. I seek to show that there is a physicalist construal of one very important and central such function, but this is not the place to attempt a complete survey of modes of explanation.

3. Irreducible Properties in Regularity Explanation

The explanation of high-level laws and regularities is, I suggest, the best place to find high-level properties doing explanatory work. It is difficult to avoid the conclusion that, for example, ravenhood will figure somewhere in the explanation of the fact that all ravens are black. Supposing—as for the sake of the argument I will—that ravenhood is irreducible, you therefore have a *prima facie* challenge for explanatory physicalism: how to explain why all ravens are black without mentioning the physically undefinable property of ravenhood? More generally, how to explain why all *F*s are *G* without mentioning *F*-ness?

The simple case of raven blackness—call it the raven explanandum—will be sufficient for my purposes in most of what follows. Some remarks on this example before I continue. First, if it is to be taken seriously, the specification of the raven explanandum must contain certain qualifications: “In *the usual environmental conditions*, all *normal* ravens are black”. But these qualifications will not in any way enhance the understanding of the explanatory role of irreducible properties, and so they will be ignored in what follows. Second, the raven explanandum is ambiguous. Is it a bare universal regularity, the simple fact of the blackness of all actual ravens, or is it something more robust or lawlike, a tendency to blackness in ravens even under certain counterfactual conditions? I will have uses for both versions of the explanandum; for the remainder of this section, however, there is no need to disambiguate.

On with the argument. I will be interested in a causal explanation of the

raven explanandum. There is a coloration mechanism that is responsible for the black color of all ravens (that is, of all normal ravens in normal circumstances). I take it that the core of any scientific explanation of raven blackness is the elucidation of this mechanism.

More specifically, I will focus on the physiology of the coloration mechanism rather than its evolutionary history. This is not to deny that there is a story about the evolution of raven blackness that complements and enriches the physiological explanation; however, the physiological part of the narrative will give me all the material I need.

What, then, is the structure of the physiological explanation of raven blackness? And what role does ravenhood play in this structure?

Let me consider three possible answers to this question:

1. Ravenhood is a direct cause of blackness: it causes blackness in a way that is unmediated by physical-level causation.
2. Ravenhood is explanatorily relevant to raven blackness because of its connection to a physical-level cause of blackness.
3. Ravenhood is not explanatorily relevant to blackness: the explanation of blackness is conducted entirely at the physical level.

Ravenhood Is Directly Causally Relevant The view that ravenhood stands in a direct causal relationship to blackness will cause great difficulty for physicalism: it is hard to see how the explanatory causal relation between ravenhood and blackness can be represented in physical language if one of its relata is physically undefinable. Or to put it in more metaphysical terms, it is hard to see how a causal relation can be purely physical when what is doing the causing is non-physical.

In the remainder of this paper I will simply assume on behalf of explanatory physicalism that ravenhood and other undefinable properties do not themselves directly cause anything: what does the actual causal work in the

explanation of raven blackness is rather an underlying physical property of ravens. If this assumption is incorrect then the reconciliation of physicalism and irreducibility presented here can only be part of the story.

I do not make this assumption out of desperation, however: if you look to some elementary raven science, you will see that the causal explanation of blackness does indeed invoke physical-level properties only. The blackness of a raven's feathers is due to a causal process in which melanosomes, cellular organelles that are produced by cells called melanocytes, first synthesize the black pigment melanin, and then migrate to outer skin and feather cells, giving them their color. The causal story about a raven's blackness, then, will make no essential reference to its ravenhood, but will rather concern such things as the expression of the gene for tyrosinase, an enzyme that catalyzes the production of melanin from the amino acid tyrosine, and about the process of phagocytosis by which the melanosomes are incorporated into the skin and feather cells.

Gather these causally relevant properties together into a property complex; call it *P*. In the light of the science, the direct causal relevance theorist must concede that *P* is a cause of blackness in ravens; in order to maintain their position, they must therefore hold that ravenhood is somehow also simultaneously a cause of blackness. The difficulties of taking such a line are well known from the literature on the exclusion problem (Kim 1998); I do not want to discuss that problem here, however, but merely to show that the explanatory physicalist's refusal to take the direct causal power of ravenhood seriously is no greater an embarrassment than the alternative.

Ravenhood's Relevance Is Mediated by Physical-Level Causes On this rightly most popular view, ravenhood is explanatorily relevant to blackness because of its relationship to the direct physical-level cause of blackness, that is, because of its relationship to the property *P* characterized immediately

above (Block 1990; Kim 1998).¹ Block, for example, writes that a high-level property is explanatory because it “brings in” a causally efficacious physical property (p. 163).² Ravenhood, then, is explanatory because it “brings in” *P*—but what is “bringing in”?

It is not easy to provide a general answer to this question. The case at hand provides an especially good illustration of the difficulties, because ravenhood and *P*-hood are quite distinct properties. Indeed, having *P* is neither necessary nor sufficient for being a raven. It is not sufficient because it is possible (indeed, quite likely) that the machinery determining raven blackness is identical in all important respects to the machinery determining the blackness of other species in the genus *Corvus*, for example, the carrion crow—thus that carrion crows as well as ravens have *P*. It is not necessary because, if the course of raven evolution had gone differently, ravens might have been brown rather than black.

If not necessity or sufficiency, then what? And why does the relationship, whatever it is, confer explanatory relevance? These questions—which I take to be the modern version of the old question in the reductionist literature about the nature of “bridge principles”—are the major issues for the mediated relevance view; they will be the subject of much of what is to follow. Before I turn to the topic of bridge principles, however, let me consider an attempt to evade the problem of high/low explanatory relations altogether.

Ravenhood Is Irrelevant The causal work in the explanation of raven blackness is to be done by the physical-level property *P*. Is anything else needed? Might it not be that the entire explanation of the blackness of ravens should be conducted in physical-level terms, without so much as using the word

1. Some writers would ascribe to ravenhood a kind of auxiliary causal relevance in virtue of its explanatory relation to *P*; for clarity’s sake, however, I will continue to talk about causal relevance only where there is a direct causal relation.

2. Block and many other writers who espouse the mediated relevance view are principally concerned with functionally defined properties, but functionalism plays no essential role in the basic proposal.

“raven”? If so, there is no role for ravenhood to play in the explanation of raven blackness.

I associate this view with what has been called the “new reductionism”, some proponents of which claim to be able to dispense altogether with principles relating high-level and physical-level properties—that is, with bridge principles (Craver 2007; Gillett 2007).

It is hard to see, however, how the no-relevance position can be sustained. To be sure, *something* can be explained in purely physical terms, namely, why all things with *P* are black, but in response to such an explanation, it is surely reasonable to say: “I understand now why everything with *P* is black, but why are all *ravens* black?” To answer this question, it is necessary to supplement the explanation with a further posit bridging the gap between ravenhood and *P*-hood: “All ravens have *P*”. With this supplement, ravenhood has entered the explanation.

Can it be argued that *All ravens are black* and *Everything with P is black* are the very same explanandum? No, because *P*-hood is neither necessary nor sufficient for ravenhood. Indeed, while the latter generalization is physically necessary, the former is contingent, since it is physically possible for ravens to evolve some other coloration (Beatty 1995; Waters 1998).

4. The Mediated Relevance Approach

I have put aside the thesis that ravenhood is directly causally relevant to raven blackness, and I have rejected the thesis that ravenhood is explanatorily irrelevant to raven blackness. In the rest of this paper, I will attempt to flesh out the mediated approach to the relevance of ravenhood.

I will suppose that the correct physiological explanation of raven blackness fits the following deductive schema, where *P* is the blackness-causing physical property complex introduced earlier and *Q* is an equally physical property that realizes blackness:

1. All ravens have P ,
 2. Having P causes Q , and
 3. Having Q is sufficient for blackness,
- ∴ All ravens are black.

The schema is familiar: it is more or less the same as that proposed by the received view of the scientific explanation of laws, and more particularly, to the schema of classical reduction (Nagel 1979). It has, that is to say, three parts: (1) a “downward” bridge principle relating the explanandum’s high-level antecedent (ravenhood) to a physical property P , (2) a causal law connecting purely physical properties, P and Q , and (3) an “upward” bridge principle connecting Q and the explanandum’s high-level consequent (blackness).

The putatively irreducible property, ravenhood, enters into the explanation only in the first of these three parts, the downward bridge principle, so it is the first part that will be my almost exclusive concern in what follows. Concerning the second part, what matters given my aims in this paper is that P ’s causing Q is derivable from the laws of physics alone. There is no need, for example, to ask here whether it is sufficient for explanatory purposes to cite a physical law “ P causes Q ”, or whether rather the mechanism underlying the law must be spelled out. Provided that the correct answers to such questions introduce only purely physical elements into the explanation—and I assume that this is the case—the identity of the elements is not important. Concerning the third part, let me just say that Q should be a physical realization of blackness.³

I call the kind of downward bridge principle that constitutes the first part of the explanation of the raven generalization a *basing generalization*.

3. For more on upward bridge principles, and in particular the differences between the constraints on downward and upward rules, see Strevens (2009, §7.52).

I assume that the form of a basing generalization is always *All Fs have P*. (Perhaps it should be statistical for statistical explananda, but let me not worry about that here.)

Now we are back at the point where I left off the discussion of the mediated relevance view in the previous section, that is, the point at which I asked: What does the downward bridge principle, the basing generalization, contribute to the explanation of raven blackness? How does the property of ravenhood help the basing generalization to make this contribution? And can this explanatory function of ravenhood be replicated by purely physical properties?

Bridge principles, and in particular the downward rules that I call basing generalizations, have been on the philosophical agenda for decades; almost all of the resulting literature supposes that they deal in necessities or sufficiencies or both. Yet no such view can provide an adequate account of the explanatory force of the raven basing generalization, since as noted above, whatever the connection between ravenhood and *P* may be, it is neither one of necessity nor one of sufficiency—whether the modality is of the logical, metaphysical, or physical variety. To provide some evidence for this claim, let me briefly consider several existing theories of explanatory bridging.

The classical view, as it is usually understood, allocates to bridge principles the task of providing metaphysically necessary and sufficient conditions; *All ravens have P* qualifies as appropriately explanatory on this view only if *P*-hood is necessary and sufficient for ravenhood. It is neither, yet the basing generalization is explanatory all the same. Thus the classical view cannot be correct.

Other views ask of a downward bridge principle only that it provide a necessary condition; such a position is inherent, for example, in the notion of explanatory relevance that Jackson and Pettit (1990) call *programming*. A property *F* programs for a property *P* if the realization of *F* ensures the realization of *P* (to use Jackson and Pettit's terminology, which is not further

elucidated). If F programs for P , then, P is necessarily instantiated wherever F is instantiated, thus the presence of P is a necessary condition for the presence of F .⁴ Perhaps Jackson and Pettit’s “ensuring” entails metaphysical necessity, or perhaps only physical necessity. Either way, ravenhood does not program for P , because it was (and perhaps still is) physically possible for ravens to evolve some coloration other than black.

A final possibility is to understand a bridge principle as stating a relationship of sufficiency; on this view, the instantiation of the physical property identified by such a principle is sufficient for the instantiation of the high-level property, which is to say more or less that the physical-level property must be a way of realizing the high-level property. Such a position is familiar from the literature on multiple realizability, as in Kim’s (1998) “functional reduction” or Melnyk (2003). The idea that the high-level and physical-level properties in a bridge principle stand in the relation of determinable to determinate is a more specific version of the same. It might occur to you that within a deductivist framework, the sufficiency view can make no sense of a *downward* bridge principle, which requires a logical move from the high level to the physical level. True enough; the view requires an amendment to deductivism. But there is a more immediate concern: the raven basing generalization does not state a sufficient condition for ravenhood, since creatures other than ravens might have (or might have had) P . To put it another way, being P is not a way of being a raven. Sufficiency cannot, then, be a condition on explanatory bridge principles.

To sum up, the rather humble example of raven blackness establishes the inadequacy of any theory of bridge principles on which they are required

4. This is how programming works when the high-level property is not multiply realizable. In cases of high-level multiple realizability, Jackson and Pettit hold that it is sufficient for programming that F ensures the realization of some property with relevantly similar causal powers to P . Thus it is not the presence of P itself, but the presence of one of a family of physical properties with the same causal profile as P , that is necessary for F -ness. This view is subject to the same objection as the simplified version presented in the main text.

to state metaphysical or physical necessities or sufficiencies. Of course, it may well be that for some basing generalizations of the form *All Fs have P*, having *P* is necessary or sufficient for being an *F*. But since not all basing generalizations are undergirded by necessities and sufficiencies, expressing such modalities cannot be an adequacy condition on basing generalizations.

What other kinds of relations could the raven basing generalization express? Not, if physicalism is to be sustained, a causal relation between ravenhood and *P*-hood, for the reason given at the end of section 3: a physicalist cannot allow that ravenhood is a direct cause of anything.

What is needed is a new picture of the relation between the high level and the physical level that makes for an explanatory bridge principle. In what follows, I will build up this picture in three steps. First, in section 5, I ask what relation between ravenhood and *P* must hold as an absolute minimum for the explanation of the bare fact of actual raven blackness to succeed. Second, in section 6, I examine a richer explanandum, a regularity with a modal aspect, and ask what needs to be added to the basing generalization for the explanation of this new kind of phenomenon to go through. Third, in section 7, I complete the picture by putting a certain relevance constraint on the facts cited by a basing generalization. The result is a theory of basing generalizations, thus of downward bridge principles, that entirely departs from the necessity/sufficiency paradigm. It is also a theory on which most basing generalizations are empirical rather than semantic or metaphysical truths—empirical truths that can be captured in a purely physical language, thus making room for explanatory physicalism.

5. Explaining Bare Regularities

Consider as an explanandum the blackness of all *actual* ravens—that is, the fact represented by Hempel’s interpretation of *All ravens are black* as a simple universal generalization about the actual color of actual ravens. I call this toy

explanandum the *raven generalization*. Once I have used it to frame the issues and to illustrate my broad strategy for dealing with irreducible properties, I will move on to the meatier case of robust regularities and laws.

I will continue to work in the deductivist framework introduced in the previous section. Suppose, then, that at the very least, the propositions that figure in the explanation of raven blackness must entail the explanandum—that propositions (1)–(3) above must entail that all actual ravens are black. (This explanatory deductivism is, like any other claim about explanation, deniable, but I have to start somewhere.) In order for the basing generalization to play its part in entailing the raven generalization, it need only state a bare empirical fact, the actual *P*-hood of all actual ravens—something far weaker than is required on the necessitarian or classical accounts of bridge principles. Assume as a working hypothesis that this exhausts its explanatory role. (I remind you that this is just a first step; there is much more to come.)

Call the explanatory state of affairs specified by a basing generalization—in the case at hand, the *P*-hood of all actual ravens—a *basing pattern*. A basing pattern, then, is the fact of a certain set of objects instantiating a certain property. It might be treated formally as an ordered pair of a set and a property. The basing pattern for the raven law, for example, would be the pair made up of the set of actual ravens—the extension of ravenhood—and the property *P*; such a pair represents the fact of all members of the set having *P*.

On this view of the explanatory significance of basing generalizations, the explanatory role of ravenhood in the explanation of the raven generalization is, obviously, to pick out the set featured in the basing pattern, that is, the set of actual ravens. But the same set could just as well be determined by some other property; a basing generalization that used this other property instead of ravenhood to pick out the ravens would pick out exactly the same basing pattern, and so make the same contribution to an explanation of the raven generalization.

Enter explanatory physicalism: you can capture physically the explanatory

contribution of ravenhood in this particular case by finding a condition framed in purely physical terms that picks out all and only the actual ravens—that has the same extension, is instantiated by the same objects in the actual world, as ravenhood.

Defining ravenhood physically may be impossible (and indeed is impossible, if ravenhood is irreducible). Finding physical conditions that are necessary and sufficient for ravenhood is at the very least extremely difficult. Even finding necessary conditions for ravenhood that have real predictive power is, given the contingencies of the evolutionary process, not so easy. But finding a physical property that is extensionally equivalent to ravenhood is by contrast rather straightforward. You need only find some property (or rather, some property complex) that all and only the actual ravens have. The property may be quite accidental, in the sense that it may be quite contingent that the ravens have the property. Perhaps it will be some kind of genetic signature, coupled with an appropriate causal relation to the rest of life on earth (to rule out raven doppelgangers from outer space). Even simpler and more foolproof, it might simply be a start point and end point on a particular branch of the actual evolutionary history of life on our planet, picking out a set containing everything that belongs to the lineage between the two points.⁵ It is hard to imagine a more accidental, contingent property of the raven species than this, depending as it does on the particular moment in history at which the raven lineage diverged from other species of the *Corvus* genus, yet such a property is quite capable of picking out the set of actual ravens—which is all it need do in order to replicate the explanatory work performed by the irreducible property of ravenhood.⁶

5. There are complications: species may occasionally interbreed, for example.

6. Can the vagueness that is perhaps implicit in the property of ravenhood be replicated physically? It does not need to be. Suppose that there are borderline cases of ravenhood. For each raven in this categorical penumbra, I ask you: must its blackness be explained in order to explain the raven generalization? If the answer is yes, then forge your physical substitute for ravenhood so that it includes the specimen; otherwise, not. If the answer to the question whether the specimen's blackness need be explained is itself vague, then I attribute

With this small example in hand, let me look forward to a grand thesis: that whatever explanatory work is done by irreducible properties, is attributable not to their causal powers but to their extensions. Call this the doctrine of the *explanatory extensionalism of irreducible properties*, or *explanatory extensionalism* for short. If extensionalism is correct, then explanatory physicalism is saved: although irreducible properties cannot themselves be captured physically, extensionalism implies that what it is about them that is explanatory, namely, their extensions, can nevertheless, at least in principle, be delineated in physical terms.

The examination, in the next section, of regularities that are more robust than the Hempelian raven generalization will show that extensionalism in its simple form cannot be maintained; a more sophisticated version of extensionalism will emerge, however, that is both correct—at least for the kinds of explananda examined in this paper—and compatible with explanatory physicalism.

6. Explaining Robust Regularities

Ravenhood need contribute just its extension to the explanation of the raven generalization, you might think, only because that generalization has no modal dimension. If the explanandum in question were not a bare empirical regularity but something more robust, a law of nature even, then its explanation would require an equally robust basing generalization; ravenhood's contribution to such a generalization would have to be far more substantial than its mere extension—quite likely substantial enough to escape the grasp of physics.

To investigate this possibility, let me develop an account of the explanation of robust regularities, based on the account of the explanation of bare

a deficiency to the explanandum, or if you like to the explanatory request, not to the physical explainer.

regularities proposed in the previous section.

You need not stray from ornithology to find a suitably robust regularity: although *All ravens are black* was interpreted by Hempel as stating a bare regularity, the phenomenon of raven blackness has a degree of robustness sufficient for my purposes. What is the nature of that robustness? It is articulated by certain counterfactual conditionals, conditionals concerning the color of ravens in counterfactual circumstances. Suppose that some particular, typical raven had hatched a day later, or had been raised on a somewhat different diet—what color would it have been? It would still have been black. Or suppose that two ravens that never in fact meet, had met and mated and successfully produced offspring. What color would those offspring have been? Black. It is in these counterfactual conditional truths that the robustness of raven blackness subsists; call them the robustness-conferring counterfactuals.⁷

The statement *All ravens are black* can be read in two ways, then: as expressing a bare Hempelian regularity, or as expressing a regularity with more modal scope. I have called the former interpretation the raven generalization; let me call the latter the *raven law*. Perhaps “law” is too strong a term, but never mind; in this paper, I attach no special importance to the term, using it only for convenience. Note that, as with the raven generalization, the raven law might better be expressed with various hedges: *In such-and-such conditions, all normal ravens are black*. For simplicity’s sake, I continue to ignore such delicate matters.

The raven generalization and the raven law state (or if you like, are) different regularities; they therefore constitute different explananda and so demand different—though of course related—explanations.

To explain the raven law, I take it, you must explain not only why all actual ravens are black, but why raven blackness is robust—why the robustness-

7. It is of course the truthmakers for the counterfactual conditionals that confer robustness; the conditionals themselves are merely linguistic constructs.

conferring counterfactuals are true. The explanation proposed for the raven generalization will therefore fall short of explaining the law: it explains why actual ravens are black in the actual circumstances, but not why ravens in counterfactual circumstances (such as actual ravens on a counterfactual diet), and counterfactual ravens (such as ravens produced by counterfactual matings), are black.

There is a simple fix: use the same explanation schema, but a more robust basing generalization. More specifically, rather than invoking the bare fact of all actual ravens' having *P*, use the same words to invoke a robust generalization: *All ravens have P*, interpreted so as to attribute *P* to (some) counterfactual ravens and to actual ravens in (some) counterfactual circumstances. (If you find the sudden importance of non-actual ravens to be ontologically disconcerting, let me assure you that talking about counterfactual ravens will turn out to be an indirect way of talking about certain features of the actual world, as explained at the end of this section. The modal metaphysics need not, then, be taken seriously in order to appreciate what I have to say.)

In which counterfactual circumstances and for which counterfactual ravens must the basing generalization hold true, if it is to serve its explanatory purpose? That is not a question that can be given a general answer; it all depends on what is being explained. The more robust is the regularity to be explained, the wider the counterfactual scope required of the basing generalization. But a basing generalization never need have wider scope than the regularity it explains. For example, there are scenarios or possible worlds where ravens evolved a brown rather than a black coloration, but to explain the raven law, a basing generalization does not have to (falsely) ascribe *P* to the ravens in such worlds, because such ravens are irrelevant to the robustness of the raven generalization—they play no role in grounding the relevant robustness-conferring counterfactuals.

The explanation of the raven law I am proposing, then, takes the form of

the following deductive argument.

1. All actual ravens (even in some counterfactual circumstances) have *P*,
as do some counterfactual ravens,
 2. Having *P* causes *Q*, and
 3. Having *Q* is sufficient for (indeed, realizes) blackness,
- ∴ All actual ravens (even in some counterfactual circumstances) are black,
as are some counterfactual ravens.

I assume for now that such a demonstration of the grounds of the robustness of raven blackness constitutes an adequate explanation of the raven law; some doubts will be addressed later in this section (“Objection Two” below).

I will now show you that this explanation is compatible with a more sophisticated form of explanatory extensionalism. The basing generalization that figures in the explanation of the raven law is more robust than the basing generalization that figures in the explanation of the bare raven generalization, but it may be understood as making the same sort of explanatory contribution, namely, a basing pattern, the fact that a certain set of objects has a certain property *P*. In the case of the raven generalization, the set is the extension of ravenhood in the actual world, or in other words, the set of all actual ravens. In the case of the raven law, the set is rather larger: it includes some non-actual ravens and some actual ravens in non-actual worlds.⁸ Loosely speaking, the larger set is the extension of ravenhood in not only the actual world but in certain nearby possible worlds, namely, the worlds in virtue of which the raven law enjoys the robustness that it does, that

8. I am counting as different objects a raven in the actual world and what we would naturally call the “same raven” in another possible world. This is for simplicity’s sake: we may want to attribute *P* to one such individual but not to the other. Counterpart theory supplies a metaphysics that makes the distinction real, but it is hardly necessary here; my way of talking can be translated into any modal metaphysics you like—including the most deflationary varieties.

is, the worlds in virtue of which the robustness-conferring counterfactual conditionals hold. Call these the raven law's *basing worlds*, and call the set of all ravens in the basing worlds ravenhood's *basing extension*.⁹ I remind you that the identity of the basing worlds, and thus the composition of the basing extension, varies with the robustness of the regularity to be explained; it is therefore relative to the explanandum. There is no such thing as *the* basing extension of ravenhood, only the basing extension of ravenhood with respect to a given explanandum, such as the raven law.¹⁰

Now it is possible to state my conclusion: the contribution made to the raven law's explanation by its basing generalization is the fact of all members of the basing extension having *P*. It follows that ravenhood's contribution to the explanation is the basing extension itself. If the same set can be captured in purely physical terms, ravenhood's explanatory contribution can be captured physically, as explanatory physicalism demands.

In short, then, the fact that the irreducible property of ravenhood has explanatory power can be seen to be consistent with explanatory physicalism by way of the following two steps:

1. The raven basing generalization's contribution to the explanation of the raven law is exhausted by its corresponding basing pattern, that is, by the fact that it implies that all objects in the basing extension have *P*, and

9. A property's basing extension, then, is poised partway between its extension in the usual sense, that is, the set of entities instantiating the property in the actual world, and the property's intension in the modern sense, that is, a function from possible worlds to sets of entities instantiating the property in a given world. Because the basing worlds are such a small subset of the possible worlds (and because it is a simple set, not a function), a basing extension is much closer to an extension than it is to an intension.

10. It is conceivable that, while some ravens in a possible world are relevant to the robustness-conferring counterfactuals, others in the very same world are not. The basing extension would then include the former but not the latter ravens. The notion of a basing world should therefore be regarded as an expository convenience; the fundamental notion is that of a basing extension.

2. The contents of the basing extension (and thus the fact of their having the physical property *P*) can be captured in physical terms.

I will soon argue for (2); first, however, I want to defend (1) against several objections, each of which proposes that the explanation of the raven law above is missing some essential part, and suggests that ravenhood makes some contribution to the missing part quite aside from its determination of the basing extension.

Objection One: Explaining the Basing Pattern In order to explain the raven law, you might think, it is not enough to state that the basing pattern obtains; you must *explain why* the basing pattern obtains. You might further suspect that the property of ravenhood—not merely its basing extension—will play a role in that explanation.

It is an interesting question whether an explanation that cites a basing generalization is complete in some sense, or whether it must provide reasons why the basing generalization is true. The answer, I think, is that the basing generalization need not itself be explained, any more than an event explanation that cites initial conditions need provide an explanation of those conditions. But there is no need to resolve this issue here. Although I hold that the explanation of the raven law proposed above is quite adequate, I am happy to concede that explaining the basing generalization cited by the explanation will cast further explanatory light on the raven law. This irresistible observation is enough to motivate the objection I have in mind. An explanation of the raven basing generalization would enhance our understanding the raven law. If ravenhood makes a contribution to such an explanation that goes beyond its basing extension, then explanatory extensionalism—now understood as the doctrine that any irreducible property’s explanatory significance is exhausted by its basing extension—is false.

The question, then, is whether ravenhood plays a role in the explanation of *All ravens have P* that is different in some way, that goes beyond, its role

in the above explanation of *All ravens are black*. The answer would seem to be no; the law and the (robustly construed) basing generalization are very much of a piece, attributing as they do certain physiological properties to ravens. You would therefore expect them to have the same sort of explanation: the basing generalization will be explained by causal mechanisms and further basing generalizations. Ravenhood will play no role in the causal mechanisms, and where it appears in the basing generalizations, it will be its basing extension that matters. To the extent that ravenhood helps to explain the basing generalization, then, it does so in virtue of its basing extension, in accordance with explanatory extensionalism.

Somehow this response, though technically correct, does not quite get to the root of the complaint. What is wanted is an explanation for the basing generalization's *robustness*, which will in turn convey an understanding of the raven law's robustness. Until you see why the basing generalization is robust, you ought not to relinquish the suspicion that ravenhood is up to something behind the scenes.

To remove all shadow of a doubt, then, let me show you why raven blackness is robust, which is to say, why the robust-conferring raven/blackness counterfactual conditionals hold true. You will see that there is no special role in the explanation to be played by the property of ravenhood.

There are a number of philosophical accounts of the truth conditions for counterfactual conditionals. To answer the present question there is no need to choose among them, however, because they agree on an algorithm for evaluating everyday counterfactuals of the sort that articulate the robustness of raven blackness—though they differ, of course, on the ultimate grounds of the algorithm's validity. I take both the fact of the consensus and the algorithm itself from Bennett (2003), to whom I refer you for the fine print.¹¹

11. For a more thorough treatment of the foundation of the counterfactuals articulating the robustness of contingent regularities such as raven blackness, including a systematic analysis of the circumstances under which these counterfactuals fail to hold, see Strevens (2008).

According to Bennett’s algorithm, when we evaluate an everyday counterfactual, we hold the actual history of the world fixed up to a point shortly before the (counterfactual) occurrence of the antecedent. We then imagine a slight deviation from the actual world’s course of events that brings about the antecedent, what David Lewis calls a “small miracle” (Lewis 1973), and ask ourselves what the causal consequences of such a deviation will be, according to the actual world’s fundamental laws.¹²

Consider, then, a typical counterfactual articulation of the robustness of raven blackness:

Had some particular, quite ordinary raven been fed a certain, quite ordinary worm the day after it hatched, it would still have been black.

What makes it true? To evaluate the conditional, according to Bennett’s consensus algorithm, we hold the history of the world fixed up until a point shortly before the raven is counterfactually fed the worm. In so doing, we hold constant the physiological structure of the raven, and indeed, the entire evolutionary history of the raven species. Most important, we hold constant

12. In applying Bennett’s consensus algorithm, I am supposing that the robustness-conferring counterfactuals are (a) not “backtracking” counterfactuals, and (b) not the kinds of counterfactuals whose truth conditions are noticeably contextually sensitive (the paradigm of sensitivity being “If Julius Caesar had commanded the UN forces in the Korean War . . .”). These assumptions are clearly true for raven blackness; you might wonder, though, whether there are other regularities whose robustness is captured in part by backtracking or contextually sensitive counterfactuals.

Context sensitivity: If a generalization is articulated so as to point to a regularity whose robustness depends on context-sensitive counterfactuals, yet the relevant context is not specified, then the robustness of the regularity is not fully specified, which is to say that the regularity itself—the explanandum—is not fully specified. I take it that an account of explanation need only concern itself with precisely characterized explananda.

Backtracking: the non-backtracking assumption applies to those generalizations of the form *All Fs are G* that hold in virtue of a causal relation running from *F*-ness, or something about *F*s, to *G*. It does not apply to, say, a generalization of the same form that holds because *F* and *G* are effects of a common cause; the robustness-conferring counterfactuals in this latter case will backtrack to the occurrence of the common cause—or so it seems to me. The analysis presented in this paper will, therefore, not apply to such explananda. But a similar analysis will apply; the philosophical lessons will remain the same.

the fact that the barely hatched raven has the blackness-producing property *P*. We then consider slight deviations from actuality that bring about the worm-feeding. In one such deviation, perhaps, a magnanimous worm-gatherer takes the left rather than the right fork in the road, and passing by the nest, tosses the hatchling a sample of his merchandise. Such a deviation, and others that are equally conservative, will not affect the young raven's *P*-hood. When we ask what, according to the actual laws, will happen next, the answer is therefore: as in any bird with *P*, the development of blackness.

The robustness of raven blackness in this case is therefore due to the following facts:

1. The hatchling actually had *P* before the putative time of the counterfactual worm-feeding, and
2. The hatchling's *P*-hood persists in the face of any conservative deviation from actuality sufficient to bring about the putative worm-feeding.

The first of these is the bare empirical fact that the raven had *P* at a certain time. The second consists in a kind of minimal resistance of that instantiation of *P*-hood to certain causal processes (in this case, due to not much more than *P*-hood's physical inertia, that is, its tendency to stay around rather than to spontaneously evaporate).

Observe how slender a foundation is sufficient to support the robustness of raven blackness. The bare fact that actual ravens have *P* is nearly enough; it gets you fact (1) above. All that needs to be added is that *P* will tend to endure if it is not actively interfered with (fact (2) above). Thus in order to transform the explanation of the raven generalization—the bare empirical fact of the blackness of all ravens—into an explanation of the robust raven law, you need only add this fact about the persistence of *P*. That is sufficient (given the modal metaphysics or semantics of robustness) to explain the additional implications of the raven law.

In other words, what needs to be added to the explanation of the raven

generalization to obtain an explanation of the corresponding law, and of the law's robustness in particular, is not a fact about ravenhood, but a purely physical fact about *P*—a fact. Explaining the raven law therefore requires no more of the unphysical property ravenhood than does explaining the raven generalization. Which is to say, it requires only ravenhood's extension—or in this case, its basing extension.

Before I consider the prospects for capturing the basing extension physically, two further worries about the extension's explanatory adequacy.

Objection Two: Rich Understanding of Laws I assumed earlier that the raven law can be satisfactorily explained by a certain basing pattern and causal mechanism. What if this assumption is incorrect? What if a statement of the raven law asserts some proposition that goes beyond what is implied by the raven basing generalization and causal mechanism? And what if the explanation of this additional content calls on ravenhood to do work that cannot be done by its basing extension alone?

The worry might be developed as follows. According to one influential account of laws of nature (Armstrong 1983), a law of the form *All Fs are G* typically asserts the existence of a necessitation relation between the properties *F* and *G*. In order to explain the law, you might think, you must explain the existence of the necessitation relation.

Suppose, to fix on the nearest available example, that *All ravens are black* is such a law. There is a necessitation relation, then, between ravenhood and blackness, and to understand the law, you must understand why the relation exists. How to do this physically, without some physical handle on ravenhood itself?

I am quite skeptical that the Armstrong doctrine applies to the raven generalization and other high-level laws with physically undefinable antecedents (and for all I know, Armstrong is equally skeptical). After all, the putative relation between ravenhood and blackness cannot be causal, or so I have assumed

on behalf of the physicalist (section 3)—not just for strategic reasons, but also because science has yet to uncover any clearly physically irreducible causal powers. Nor can it be necessitation in any of the usual senses, since ravens are not necessarily black: there is nothing about ravenhood that precludes the possibility, noted above, of ravens evolving some other coloration (as other species in the genus *Corvus* have in fact done). Until some positive suggestion is made as to the identity of the necessitation relation—no doubt with the help of an example other than raven blackness—I will put this objection to one side.

Objection Three: Explaining Extension A third complaint is that, in order to understand the raven law, you need to understand how ravenhood gets its basing extension, or in other words, that you need to understand in virtue of what ravenhood picks out the objects that it does in the actual and other basing worlds. (This complaint would apply also to my proposed explanation of the bare raven generalization.) Clearly, something about ravenhood other than its basing extension is going to have to appear in such an explanation.

The objection is, I believe, fundamentally misconceived. Were the line of thought it implicitly endorses to be taken seriously, we would have to say that science does not fully understand the behavior of, say, electrons until it has an explanation of why the word *electron* refers to electrons, or at least, of why the property of being an electron has the extension that it does. While such an explanation would be of great philosophical interest, its formulation is hardly a necessary condition for fruitful physics—indeed, it is not a proper part of physics at all.

Physically Capturing the Basing Pattern Suppose that the foregoing objections have been met satisfactorily, and that the raven basing generalization's contribution to the explanation of blackness is indeed nothing over and above a basing pattern, that is, a pattern of *P*-hood among the ravens in the basing worlds. Can this pattern be captured physically?

There are two approaches to finding a physical formulation that picks out the basing pattern. The direct approach focuses on the problem of finding a physical condition that picks out ravenhood's basing extension, that is, a condition that is satisfied in the basing worlds by all and only those worlds' ravens. The indirect approach focuses rather on finding a physical condition that picks out the actual-world facts in virtue of which we say that some raven in the relevant counterfactual circumstances has *P*, that is, the facts about the actual world in virtue of which the basing pattern holds.

Let me begin with the indirect approach. Here the work has already been done, because the problem of showing that the actual-world facts that *make the basing generalization true* are purely physical is not substantially different from the problem of showing that the actual-world facts that *explain the basing generalization* are purely physical—and this latter problem was tackled as a part of the response to Objection One above. I will recap briefly. The basing worlds have the following configuration: they are identical to the actual world up until a certain time, and then they deviate conservatively so as to bring about a certain state of affairs (the antecedent of the some robustness-conferring counterfactual) at a slightly later time. The robust basing generalization is true because, at this slightly later time, all ravens in a basing world have *P*. This is true in virtue of the conjunction of two actual-world facts: the *P*-hood of all actual ravens, which ensures the presence of *P* up to the point at which the deviation occurs, and the fact that *P* survives the deviation, a matter of its persistence in the face of certain causal processes (such as the consumption of a worm). These two kinds of facts about the actual world, then, constitute the metaphysical foundation of the robust basing generalization. As shown above, they can be captured physically. The actual *P*-hood of all actual ravens can be physically specified provided that an expression can be formulated in a purely physical vocabulary that has the same extension as ravenhood; I have argued that such expressions are available. As to the persistence of *P* in the face of various conservative

(or other) deviations, that is a fact, however complex, about the physical property P , and so can also be physically specified. In short: the facts about the actual world in virtue of which the robust basing generalization holds can be specified physically, thus the facts in virtue of which the generalization explains the raven law can be specified physically, as desired.

The indirect approach should satisfy you—but just in case, let me consider the direct approach to capturing the robust generalization, which turns on the question whether the basing extension of ravenhood (relative to the explanandum, the raven law) can be picked out by a physical expression. Is there such an expression? Observe that the basing worlds share the same biological structure as the actual worlds, and in particular, that ravens in the basing worlds have the same evolutionary history and the same physiology as actual ravens (since it is in virtue of the shared biology that the raven law enjoys the robustness that it does). It follows that the variation in ravenkind in the basing worlds will not be significantly greater than the variation in the actual world: the same biological facts constrain the possibilities in each case. An expression that captures the actual extension of ravenhood in a reasonable way will for this reason quite likely capture the basing extension of ravenhood without any reworking at all.

(What is it to capture the actual extension of ravenhood in an “unreasonable” way? It is to capture it by using some unbiological contingency, for example, by way of the expression *Poe’s iconic birds*. Such an expression will not capture the basing extension of ravenhood because it does not robustly pick out ravens under all the counterfactual suppositions relative to which raven blackness is robust. To put it another way, the corresponding basing generalization, *Poe’s iconic birds have P* , will not have the same degree of robustness as the standard basing generalization, because unlike the standard basing generalization, it will not imply the P -hood of ravens under counterfactual suppositions about Poe’s literary output. Certain related attempts to capture basing extensions in quirky ways will be ruled out for different

reasons in section 7.)

Whether you consider the direct or the indirect argument, what is important is that the baseline for evaluating all counterfactuals relevant to the robustness of raven physiology is the actual physiology of the raven species. Even a claim such as “The raven species might have been brown” is true in virtue of what might have happened to the raven species (or its precursors) in the light of its actual physiology at some early time—a physiology that can, I assume, be physically specified.

7. A Relevance Constraint

The account of regularity explanation offered in the previous two sections is, I believe, incomplete. Further, it is incomplete in a way that suggests that there is an additional dimension to the role played by ravenhood in the explanation of the raven law—that ravenhood has a kind of explanatory relevance to blackness that has not yet been explored. In what follows, I investigate this additional component of relevance, and show that it does not undermine explanatory extensionalism.

Following Kyburg (1965), say that a raven is *hexed* if it is the object of a certain magic ritual. There is no such thing as magic, of course, but the ritual is real—raven hexers really do enunciate certain portentous formulas while gesticulating mysteriously at their ravens, even if the words and gestures have no effect—and thus we may truthfully say that certain ravens are hexed, meaning only that they have been subject to this empty rite.

Suppose that you want to explain the fact that all hexed ravens are black. This is an entirely legitimate explanatory goal—hexed ravens are black, and there is a reason why. What is this reason? How should the explanation of blackness in hexed ravens proceed?

If you unreflectively adapt the explanation of the raven law above, your explanation of hexed ravens’ blackness will have the following form:

1. All hexed ravens have P ,
2. Having P causes Q , and
3. Having Q is sufficient for blackness.

Parts (2) and (3) are of course identical to the corresponding parts of the explanation of the raven law; only the basing generalization (1) has changed. It will be my topic in this section.

Observe that (1) satisfies all of the requirements placed on explanatory basing generalizations in the previous section. The explanation of hexed-raven blackness therefore appears to be entirely in order. From the fact that ravenhood contributed its basing extension to the basing pattern explaining raven blackness, I concluded that ravenhood is explanatorily relevant to raven blackness. It would seem, then, that from the fact that *hexed* ravenhood contributes its basing extension to the corresponding basing pattern explaining hexed-raven blackness, I should conclude that hexed ravenhood is explanatorily relevant to hexed-raven blackness.

But this, as Henry Kyburg and other writers pointed out long ago, is unacceptable. It is fine to say that ravenhood is relevant to the explanation of blackness in hexed ravens, and not fine at all to say that hexing is relevant. Ravenhood has a place in the explanation of blackness that hexing does not. Until we understand the difference in explanatory potential between ravenhood and hexing, we have not fully understood ravenhood's explanatory role even in the limited case of blackness.

(Note that this feature of our explanatory practice constitutes a further objection to the necessitarian understanding of bridge principles that I have attributed to Jackson and Pettit, since if the instantiation of F necessitates the presence of—or “ensures the realization of”— P then a fortiori, the instantiation of “hexed F -ness” also necessitates the presence of P , rendering hexing, when present, explanatorily relevant to absolutely everything.)

The question, then: what is wrong with invoking the property of hexing to explain blackness?

There is a connection between the kind of relevance that hexing evidently lacks and a certain kind of counterfactual dependence: all hexed ravens may be black, but even if they were not hexed, they would still be black. Contrast with the case of ravenhood: if the hexed ravens had not been ravens then it is hard to say what color they would have been. They might still have been black, but depending on the respect in which they lacked ravenhood, they might well have been some other color. We cannot say, then, that if they had not been ravens, they would still have been black. Thus ravens' blackness depends counterfactually on their ravenhood, which we judge to be explanatorily relevant to blackness, but not on their hexing, which we judge to be irrelevant.

I propose to turn this observation into a test for explanatory relevance: in order to qualify for inclusion in an explanation, a basing generalization's antecedent must participate in an appropriate relation of counterfactual dependence. I will formulate the test only for cases in which a generalization of the form *All Fs are G* is to be explained in part by a basing generalization of the form *All \widehat{F} s are P*, where *P* is some appropriate *G*-causing property and \widehat{F} -ness is either a constituent of *F*-ness (so that all *F*s are by definition \widehat{F} s) or is *F* itself—if *F* is hexed ravenhood, for example, \widehat{F} might be ravenhood. The test, which determines what is and is not an appropriate choice for \widehat{F} , requires that the *F*s' *P*-hood counterfactually depend on their \widehat{F} -ness, in the sense that, if an *F* were not \widehat{F} , it might not have had *P*. (There is no interesting difference, I think, between this test and a test that requires that *F*s' *G*-ness depend on their \widehat{F} -ness, since the dependence of *G* on \widehat{F} will go by way of the dependence of *P* on \widehat{F} .)¹³

13. This is perhaps not quite the test you expected. I suggested above that, intuitively, the problem with *All hexed ravens have P* as a basing generalization is that *P*-hood (or blackness) does not counterfactually depend on hexing. According to the test I have formulated, the problem is rather that *P*-hood does not counterfactually depend on "hexed ravenhood". I

The counterfactual test for relevance constitutes a new and additional component of the account of regularity explanation sketched above. With this explanatory supplement, ravenhood is now explanatorily relevant to blackness in two ways: by way of a relation between *P*-hood and a set of ravens (the relation in virtue of which the basing pattern obtains), and by way of a relation between *P*-hood and ravenhood (the relation of counterfactual dependence in virtue of which the new relevance requirement is satisfied).

This would seem to spell deep trouble for explanatory extensionalism: ravenhood's new, second kind of explanatory relevance rests on a metaphysical dependence relation between *P* and the property of ravenhood, rather than on a relation between *P* and some extensional entity.

But extensionalism can be preserved: a property that has the same basing extension as ravenhood will pass the relevance test if and only if ravenhood itself passes the test; more generally, any two properties with the same basing extension will pass or fail the test together. The counterfactual relevance test, then, distinguishes properties only at the grain of their basing extension.

Why? Observe first that passing or failing the counterfactual test depends only on the way things are in the basing worlds. When we ask what would happen if a certain raven or group of ravens were not hexed, or what would happen if they were not ravens, we look to the closest possible worlds in which such an antecedent is realized. These will be worlds in which ravens have the same evolutionary history as in the actual world, and so they will be basing worlds, or close enough. (The significance of the "close enough" will be discussed shortly.)

Now suppose that some property *R* has the same basing extension as

take it, however, that the closest worlds in which an actual hexed raven does not have its property of hexed ravenhood are worlds in which it is an unhexed raven—or in other words, that the most conservative way to strip any bird of its hexed ravenhood is to strip it of its hexing. It is therefore not misleading to say that hexed ravenhood fails the test because *P*-hood does not depend on hexing; I will take advantage of this observation to simplify the discussion in the main text.

ravenhood. That ravenhood passes the counterfactual test (with respect to P) is due to the fact that in some closest possible worlds in which a bird is stripped of its ravenhood, it also lacks P -hood. By the argument in the previous paragraph, these worlds are basing worlds, worlds in which a bird is a raven just in case it has R . Thus in some of the closest possible worlds—the same ones—in which a bird that is actually R lacks R , it also lacks P , and so R passes the counterfactual test. Similarly, since hexing fails the counterfactual test with respect to P , any other property with the same basing extension as hexing will fail the test—because the closest possible worlds in which ravens are not hexed are also the closest possible worlds in which they lack the other property, and in this set of worlds, the ravens in question, however you describe them, all have P .

The argument is entirely general: if property P counterfactually depends (or fails to depend) on property F , and if F is extensionally equivalent to another property G in all the worlds relevant to evaluating this particular kind of counterfactual dependence, then P must also counterfactually depend (fail to depend) on property G , since the conservative deviations or “small miracles” that remove an object’s F -ness in such worlds are identical to the deviations or miracles that remove the same object’s G -ness, thus either way you are asking of the same set of objects whether they all have P .

You might nevertheless feel dissatisfied. True, the counterfactual test does not discriminate between properties with the same basing extensions, but it does not follow that the properties themselves do not play some part in explaining the counterfactual dependence. Might it not be, for example, that although P -hood counterfactually depends on both ravenhood and R -hood (where R is a physical property with the same basing extension as ravenhood), what explains this dependence is the nature of ravenhood? Might it not be, in other words, that P depends on R -hood only because it depends on ravenhood, and R -hood happens to be a good approximation to ravenhood in the local modal neighborhood? If so, ravenhood would be doing some

covert explanatory work. To put it another way, if the explanatory relevance of *R*-hood to blackness depended on the explanatory relevance of ravenhood to blackness, then substituting the former for the latter, though it might create the impression of a purely physical explanation, would do so by concealing rather than bypassing ravenhood's explanatory potency.

In order to foreclose this possibility, let me dig deeper into the reasons that ravenhood passes, but hexing fails, the counterfactual relevance test. You will see that insofar as the properties of ravenhood and hexing come into the picture, it is in virtue of their basing extensions alone.

Why, then, would a hexed raven, if it had not been hexed, still have had *P*, while the same hexed raven, if it had not been a raven, might possibly not have had *P*?

The philosophical mind has been habituated, over the last twenty or thirty years, to pass from counterfactual dependence to causation with hardly a qualm. You might therefore be inclined to explain the facts about *P*'s counterfactual dependence on ravenhood but not hexing by proposing that ravenhood but not hexing causes *P*. This, it turns out, is incorrect. Causation does come into the picture but in a more subtle way.

Consider the procedure for evaluating counterfactuals such as *If this hexed raven had not been hexed, it would still have had P*. As explained earlier, we apply the following algorithm. We restrict our attention to possible worlds that are identical to the actual world, up until a time shortly before the antecedent hypothetically comes to obtain. We narrow our focus further to the subset of these worlds in which there is a small deviation from the actual course of events that causes the antecedent to come about. Finally, we let the worlds in the subset evolve according to the actual laws of nature, to see what the causal consequences of the deviations might be.

In the case of the hexed raven that is counterfactually unhexed, then, we look at worlds that are identical to the actual world up until shortly before the raven's hexing. In such worlds the raven has *P*, just because this is the way

things are in the actual world. We then focus on the worlds within this set where hexing is prevented by a very conservative deviation from the actual course of events—worlds in which the raven enchanter’s car breaks down on the way to the hatchery, and so on. There will be many such worlds, and the hexing will be prevented slightly differently in each. In every case, however (or perhaps in all but a negligible number of cases), the deviation that prevents the hexing will not affect the raven’s *P*-hood. Why not? The physical realization of a raven’s hexing and the physical realization of its *P*-hood are sufficiently distinct that a conservative change to one will easily bypass the other.

Now take the counterfactual *If this hexed raven had not been a raven, it would still have had P*. To evaluate the counterfactual, we begin with a world in which all ravens have *P*, as always, but we must now find causal processes that, as conservatively as possible, strip the raven of its ravenhood.¹⁴ It is possible to see, without inquiring too deeply into the nature of such processes, that some of them would affect the raven’s internal workings. In these same internal workings, however, inheres a raven’s *P*-hood. Interfere with one, and you will surely—in some possible scenarios at least—disturb the other. Some “small miracles” that deprive the bird of its ravenhood will also deprive it of its *P*-hood. The resulting specimen will not (or might not) be black.

Let me put this in a form more susceptible of generalization. The property *P* counterfactually depends on ravenhood but not hexing because the physical realizer of ravenhood but not hexing overlaps (or tends to overlap) with either:

1. The realizer of *P*-hood, or

14. It is possible that when reasoning about dramatic changes such as a transformation of species, we do not require that the counterfactual antecedent be brought about by a causal process; some other kind of “small miracle” might be called upon to do the trick. The story in the main text will then have to be amended, but not, I am confident, in a way that will impact the philosophical moral.

2. The realizer of some creating or sustaining cause of *P*-hood.

Disrupting an instance of ravenhood means interfering with its realizer; this interference stands either to constitute or to cause interference with the realizer of *P*-hood, and so—at least in some cases—to disrupt the instantiation of *P* itself.

The metaphysical dependence relations that enter into this story are causal relations; ravenhood itself does none of the causing, however. In fact, ravenhood itself does nothing at all. Nor does *P*. (Though of course, it is thanks to *P*'s causal powers that *blackness* counterfactually depends on ravenhood.) What matters is not the nature of the properties themselves, but the fact that their realizers are physically or causally entwined. And this fact depends on two things: first, purely physical facts about overlap and causation, and second, the facts about the way that the properties are realized in the relevant possible worlds, or in other words, the facts about the properties' basing extensions.

In short, insofar as the properties enter into the explanation of the counterfactual dependence relations that hold between them, they do so in virtue of their basing extensions, which might just as well be called their basing realizations. It is no coincidence, then, that properties that are coextensive in the basing worlds enter into exactly the same counterfactual dependence relations.

One loose end: The relevance criterion requires a slight expansion of the notion of the raven basing extension. Previously, the basing extension included all birds relevant to the evaluation of counterfactuals of the “If this raven had eaten this worm . . .” variety. Now it must also include birds relevant to the evaluation of counterfactuals of the “If this bird had not been a raven” variety.¹⁵ The relevant birds in the latter case are ravens that define the

15. It would be neater, though fussier, to distinguish two basing extensions for ravenhood relative to any given explanandum: the extension relevant to the robustness-conferring counterfactuals and the extension relevant to the counterfactuals figuring in the relevance

outer boundaries of “nearby ravenhood”, and so determine which counterfactual tweaks most conservatively strip a bird of its ravenhood. Call these the boundary cases (no reference to vagueness is intended). A physical expression capable of capturing the new, expanded basing extension must therefore be able to determine, given ravens altered in various ways so as to take them near the boundary, whether they are still ravens. Two considerations show that such expressions are not so difficult to come by. First, the expression need be expert in only those boundary cases that result from the most conservative techniques for removing ravenhood. It may be that species membership is most conservatively removed by rendering an animal unable to breed with other members of the species (where the inability is for profound reasons rather than because of minor malfunctions). The expression, then, must attend to interbreeding, but not to other important concomitants of ravenhood. Second, the boundary cases are created by altering ravens with the actual raven biology. Thus the expression need be expert only in the physiology of the interbreeding of actual ravens. I take it that this is within the physicalist’s reach.

You are now convinced, I hope, that the counterfactual relevance test demands from ravenhood no more in the way of an explanatory contribution than its basing extension, and so sits happily with explanatory extensionalism. Capture ravenhood’s—or any other irreducible property’s—basing extension, and you have captured everything about it that is of explanatory significance.

* * *

In reaching this physicalist waypoint, I should note, I have sketched the outlines of a theory of the explanation of certain kinds of physically contingent causal regularities, and in particular, a theory of the kind of downward bridge principle that appears in such explanations, a theory according to

criterion. It is only the ravens in the former basing extension that need have the blackness-producing property *P*, whereas both extensions must be physically articulable. I hope that there will be no real confusion if, for simplicity’s sake, I conflate the two extensions.

which such rules typically state not semantic or metaphysical facts, nor even physical necessities, but rather contingent physical facts.

To summarize the view: A regularity of the form *All Fs are G* that is true not because *F*-ness itself causes *G*, but because of a contingent connection between *F*-ness and some *G*-causing property *P*, should be explained by citing three kinds of fact. First is the purely physical, purely causal fact that *P* causes *Q*; second is the metaphysical fact that *Q* is one way of realizing *G*; third—and here is the basis of the required downward bridge principle—is a basing pattern captured by a generalization of the form *All \widehat{F} s are P* (where \widehat{F} is either *F* itself or some more inclusive property).¹⁶

I have been almost entirely concerned with the question of the constraints on the basing pattern, and more particularly on the property \widehat{F} , which I claim amount to the following:

1. \widehat{F} s in all basing worlds (including the actual world) must have *P*.
2. *F*s' having *P* must counterfactually depend on their having \widehat{F} , in the sense that, had they not had \widehat{F} , they might not have had *P*.

The satisfaction of conditions (1) and (2) together implies an intimate modal connection between \widehat{F} and *P* that makes itself manifest in a certain kind of two-way conditional dependence:

1. All \widehat{F} s have *P*, and if you interfere with an \widehat{F} then provided that you do not undermine its \widehat{F} -ness, it will still have *P*. Indeed, if you interfere with the world so as to bring a new \widehat{F} into existence, it too will have *P*. You might write this: *If \widehat{F} then P* (remembering that its meaning is quite different from the meaning of the corresponding material conditional, though it entails that conditional).
2. But if you go further and undermine an existing thing's \widehat{F} -ness you will in a significant number of cases undermine its *P*-hood as well. You

16. The view is generalized to other classes of regularities in Strevens (2009, chap. 7).

might— somewhat recklessly!— write this *If \widehat{F} then $\neg P$* (remembering that it does not even entail the corresponding material conditional).

When these conditions hold, say that \widehat{F} is *entangled* with P (noting that one property's entanglement with another depends only on the properties' basing extensions).¹⁷ Entanglement comes in degrees; the more robust a regularity, the greater the degree of entanglement required to explain it.

Three observations about the scope of this account. First, in order to explain *All Fs are G* along the lines sketched here, it is not necessary that \widehat{F} be entangled with every property that participates in the causation of the F s' G -ness. It might be that \widehat{F} is entangled with some P that is sufficient to bring about G only in conjunction with certain other initial conditions Z ; in this case, the generalization to be explained is perhaps better phrased *In conditions Z, all Fs are G*, even if the conditions Z hold ubiquitously.

Second, an explanation might cite multiple basing patterns: F -ness might be constituted by two properties F_1 and F_2 , entangled respectively with physical properties P_1 and P_2 , which work together to produce G . In such a case, F_1 and F_2 have separate counterfactual tests to pass; otherwise, everything works in the same way.

Or third, the multiple basing patterns might pick out physical properties that represent not parts of a single causal route to G , but alternative routes to G . It might be, in other words, that F is sometimes entangled with P_1 , sometimes with P_2 , and that either of P_1 or P_2 is sufficient, in the context in which it typically appears, to cause G ; together, then, the entanglements explain why all F s are G . The validity of such an explanation depends on issues concerning multiple realizability that I have not discussed here. But there is nothing about my account of basing generalizations that rules out this sort of explanation; in other words, the kind of one-to-many downward

17. Strictly speaking, entanglement should be regarded as a relation between property instances rather than properties; a basing pattern may then be regarded as a pattern of instance entanglement (Strevens 2009, §7.34).

bridge principle envisaged by Fodor (1974) is quite compatible with my account of the explanatory relevance of the high level.

8. Conclusion: Relevance in the Explanation of Laws and Events

The lessons learned from raven blackness can be generalized, I propose. Irreducible—which is to say, physically ineffable—properties do not enter into explanations because of their causal relevance; as best as our science can determine, they have no such relevance. The most powerful argument for this thesis is, as I have said, empirical: wherever we find an irreducible property apparently implicated in a causal processes, close inspection shows that it is not the property itself that does the causing, but rather some physical mechanism associated with the property.

How is the irreducible property explanatorily relevant, then? It is entangled with the causally efficacious property. When a property's explanatory relevance goes by way of entanglement, it does its explanatory work in virtue of its basing generalization alone. Thus, the explanatory function of an irreducible property can be performed by a physical property with the same basing extension. Thanks to the narrowness of sets of basing worlds, such properties are, I have suggested, easy to find. It follows that, although irreducible properties do have an explanatory role to play in the high-level sciences—often a large and important role—the same role can be played by physical properties, as explanatory physicalism demands.

The theory of entanglement has been developed using the single toy example of raven blackness, but it provides a plausible understanding of the explanatory role of a wide range of high-level entities, including many of the things earlier suggested as possibly physically irreducible: psychological states, genes, chemical categories such as water, social practices, and so on. In each case, the entity, property, or category does its explanatory work, I suggest, in virtue of its local entanglement with a physical-level property. Note that a

high-level property need not be physically undefinable to participate in an explanation by way of entanglement; the problem of irreducibility serves to motivate the account of an entanglement relation, but it does not thereby restrict its explanatory scope.

What obstacles might there be to the generalization of my claims? I see two. First, there may be explanatory roles played by irreducible properties that I have not considered, and which may not be so easy to fill physically. Observe, however, that it is bridge principles that have presented the greatest philosophical challenge to explanatory physicalism, and that my account of irreducible properties' explanatory relevance is built around an explanatorily physicalist theory of the nature of (downward) bridge principles, that is, a theory of such rules on which they turn out to be, for explanatory purposes, purely physical. Though there may be other explanatory roles for irreducible properties, then, none has been taken to pose such a problem for the physicalist as the role treated here.

A second obstacle: there may be basing extensions that are not so easy to capture using the expressive resources of physics. You might worry in particular about basing generalizations that are close to physical necessities, and thus for which the basing worlds are a very broad set. This and the previous sources of concern are best handled, I think, by inviting opponents of explanatory physicalism to point to a particular explanatory role, or a particular basing extension, that jeopardizes physicalism.

Ravenhood's explanatory role emerged entirely within the context of regularity explanation. What of event explanation? There are two ways to explain, say, the blackness of a particular raven. First, you may cite particular properties of the raven in question that were causally responsible for its blackness. You may point, in other words, to the fact that the raven has property P , and then show that P causes blackness. Where x is the raven, such an explanation has the following two parts:

1. Object x has property P , and

2. Property P causes blackness.

All properties cited in such an explanation play a causal role. Ravenhood, obviously, does not appear in the explanation.

The second kind of explanation has the same causal elements as the first, but instead of citing specific initial conditions it cites x 's ravenhood along with a basing pattern:

1. Object x is a raven,
2. All ravens have property P , and
3. Property P causes blackness.

This explanation is equivalent, I believe, to explaining x 's blackness by subsuming it under the raven law, since (2) and (3) together constitute the metaphysical foundation of the raven law.

Either explanation is legitimate. The second accords ravenhood an explanatory role, namely, the very same role that it plays in the explanation of the raven generalization or the raven law: it contributes its basing extension to the basing pattern. The work done by ravenhood in event explanation, then, may for the same reasons as in the case of law explanation be replicated in a purely physical language.

Let me conclude with what you might think is an objection to this paper's principal thesis.

On the account of regularity explanation presented above, the complaint runs, the property of ravenhood may play an important part in the usual explanation of the raven law, but it is not genuinely explanatorily relevant to blackness for the following reason. Given a phenomenon to be explained, an explanation will single out a set of facts in the world that bear an objective, explanatory relation to the phenomenon. Call these the explanatory facts. For a causal account of explanation, for example, the explanatory facts will be the causes of an explanandum along with the laws in virtue of which they are

causes. But the raven law's explanatory facts do not, on my account, include the property of ravenhood. The nearest they come is their inclusion of a basing pattern, the fact of all ravens having *P*. This state of affairs, although it may be conveniently picked out using the property of ravenhood, does not, as I have been at pains to insist, essentially involve that property, but only its basing extension. Thus, although the set of all ravens (in the basing worlds) is among the explanatory facts—because the basing pattern is among the facts, and the set is a building block of the pattern—ravenhood itself is not.

There are two senses in which a fact can be an explainer. In the first sense, to be an explainer is to be among the explanatory facts. In the second sense, to be an explainer is to be used to communicate the explanatory facts. The property of ravenhood is an explainer in the second but not in the first sense. But only explainers in the first sense, you might think, play a genuinely explanatory role; explainers in the second sense play a merely linguistic (or other representational) role. They get at the explanatorily relevant facts, but they are not explanatorily relevant themselves, except in a derivative and uninteresting way.

In response to this objection, I happily concede that ravenhood is not among the explanatory facts, in the sense defined above. I have nevertheless achieved, I propose, what I set out to do: to determine the sense in which ravenhood plays a role in the explanation of blackness, and more generally, to determine the sense in which an irreducible property such as ravenhood can play a role in explanation even in a world where everything that happens, happens for physical reasons.

It turns out that the explanatory contribution made by ravenhood and other irreducible properties in the higher-level sciences is not a matter of their turning the cogs in the metaphysical engine room—but what did you expect? Our best science provides powerful reasons to think that what goes on behind that door is due to purely physical properties alone.

References

- Armstrong, D. M. (1983). *What Is a Law of Nature?* Cambridge University Press, Cambridge.
- Beatty, J. (1995). The evolutionary contingency thesis. In G. Wolters, J. G. Lennox, and P. McLaughlin (eds.), *Concepts, Theories, and Rationality in the Biological Sciences*, pp. 45–81. University of Pittsburgh Press, Pittsburgh.
- Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford University Press, Oxford.
- Block, N. (1990). Can the mind change the world? In G. Boolos (ed.), *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge University Press, Cambridge.
- Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press, Oxford.
- Fodor, J. A. (1974). Special sciences. *Synthese* 28:97–115.
- . (1997). Special sciences: Still autonomous after all these years. *Philosophical Perspectives* 11:149–163.
- Garfinkel, A. (1981). *Forms of Explanation*. Yale University Press, New Haven, CT.
- Gillett, C. (2007). Understanding the new reductionism: The metaphysics of science and compositional reduction. *Journal of Philosophy* 104:193–216.
- Jackson, F. and P. Pettit. (1990). Program explanation: A general perspective. *Analysis* 50:107–117.
- Kim, J. (1998). *Mind in a Physical World*. MIT Press, Cambridge, MA.
- Kyburg, H. E. (1965). Comment. *Philosophy of Science* 32:147–151.

- Lewis, D. (1973). *Counterfactuals*. Harvard University Press, Cambridge, MA.
- Melnyk, A. (2003). *A Physicalist Manifesto: Thoroughly Modern Materialism*. Cambridge University Press, Cambridge.
- Nagel, E. (1979). *The Structure of Science*. Hackett, Indianapolis, IN.
- Pereboom, D. and H. Kornblith. (1991). The metaphysics of irreducibility. *Philosophical Studies* 63:125–145.
- Putnam, H. (1975). Philosophy and our mental life. In *Mind, Language and Reality: Philosophical Papers*, volume 2. Cambridge University Press, Cambridge.
- Strevens, M. (2003). *Bigger than Chaos: Understanding Complexity through Probability*. Harvard University Press, Cambridge, MA.
- . (2004). The causal and unification accounts of explanation unified – causally. *Noûs* 38:154–176.
- . (2005). How are the sciences of complex systems possible? *Philosophy of Science* 72:531–556.
- . (2008). Physically contingent laws and counterfactual support. *Philosopher's Imprint* 8(8).
- . (2009). *Depth: An Account of Scientific Explanation*. Harvard University Press, Cambridge, MA.
- Waters, C. K. (1998). Causal regularities in the biological world of contingent distributions. *Biology and Philosophy* 13:5–36.