

Explanatory Autonomy and Explanatory Irreducibility

Michael Strevens

Draft of September 2010

ABSTRACT

A powerful argument for anti-reductionism turns on the premise that the biological, behavioral, and social sciences are, in the way that they explain their characteristic subject matters, in some sense autonomous from physics. The argument is formulated and strengthened in this paper, and then undermined by showing that a reductionist account of explanation is not only consistent with, but provides a compelling account of, explanatory autonomy. Two kinds of explanatory abstraction, objective and contextual, play important roles in the story.

The physical world, there is ever more reason to think, is the only world we have. Everything is made of physical stuff, and everything that happens, happens because of the way that physical laws push this physical stuff around. This is the doctrine of physicalism. It is deniable, but is in the ascendancy.

You might think, then, that a closely related doctrine that I will call *explanatory reductionism* would also command the allegiance of the vast majority of philosophers of science. Explanatory reductionism is the conjunction of two other theses. The first thesis, explanatory physicalism, says that everything about the world that can be explained, can be explained in physical terms, that is, in terms of physical facts and properties of the fundamental laws of nature. (A physical fact is any fact that can be captured using vocabulary defined using the language of fundamental physics and the usual logical connectives and other syntactic apparatus.)¹ The second thesis, which might be called explanatory monism, says that the same rules that determine what makes for a good explanation in physics also determine what makes for good explanations in the other sciences. Explanatory reductionism, then, is the view that the rules for evaluating explanatory relevance and so on in physics, together with the fundamental physical facts, determine the correct explanations of all phenomena—biological, psychological, sociological, economic.

Explanatory reductionism, though it has its supporters, is considered dubious by many philosophers. Why? The purpose of this paper is to sketch one feature of scientific explanatory practice that has been thought to count against explanatory reductionism—explanatory autonomy—and to show how certain, not especially proprietary features of Strevens’s (2009) kairetic account of explanation can reconcile autonomy with reductionism.

1. Perhaps the vocabulary of “fundamental metaphysics” should be added to the list, if it is needed to enable talk of mereological relations, causal connections, and so on.

1. From Autonomy to Irreducibility

To say that the enterprise of explanation in, say, economics is autonomous is to say that it is in some sense independent of the enterprise of explanation in physics, or that economic explanations in some sense float free of physical explanations.

Fodor (1974) both characterizes explanatory autonomy and sets it in opposition to reductive views:

Reductivism . . . flies in the face of the facts about the scientific institution: the existence of a vast and interleaved conglomerate of special scientific disciplines which often appear to proceed with only the most token acknowledgment of the constraint that their theories must turn out to be physics ‘in the long run’. (112–113)

This is to describe a methodological autonomy, but Fodor thinks that the methodological autonomy is a consequence of a something more metaphysical. Writing about economics, for example, he remarks:

What is interesting about monetary exchanges is surely not their commonalities under *physical* description. (103–104)

From these kinds of observations, Fodor concludes that there is no principled connection of explanatory interest between the properties that do work in special sciences and the properties that do work in physics, aside from the minimal connection implied by the truth of physicalism (as distinct from explanatory reductionism).

The reasoning behind this conclusion can be divided, I think, into two parts. First, explanation in physics accords relevance to small details—details of implementation—that are explanatorily irrelevant from the perspective of the higher-level sciences; further, in the cases where the special scientific kinds in question are radically multiply realizable, the irrelevant details of implementation may not even be that small.

Second, the kinds of properties that matter for explanatory purposes in the higher-level sciences are not the kinds of properties that matter in physics. Fodor does not go so far as to say that high-level explanatory properties are undefinable in physical terms. Rather, he says that they are not physically “natural”; they are not a part of the explanatory toolkit of physics.²

Fodor’s discussion is focused on high-level kinds that appear to have a functional nature—that appear to be individuated by their causal role—such as the kinds of belief/desire psychology and economics. But the same points have been made about explanation in other sciences, most particularly, biology. Kitcher (1984, 2001), writing about the explanation of Mendel’s laws and the evolution of the 1 : 1 sex ratio, and Sober (1983), writing about equilibrium explanation generally, have argued that first, physical details, even concerning quite high-level matters of implementation, frequently do not matter for explanatory purposes, and second, that what does matter is often not the sort of thing that matters in the explanations found in physics.

Bundle these two general observations together and call them the thesis of explanatory autonomy. The thesis might be stated methodologically: *The rules for constructing explanations, for determining explanatory relevance, are different in the high-level sciences than in physics.* Or it might be stated in more metaphysical terms: *What is explanatorily relevant in the high-level sciences is quite different from what is explanatorily relevant in physics.* Take your pick.

How does explanatory autonomy bear on explanatory reductionism? The two doctrines are logically consistent. First, explanatory reductionism does not entail that low-level physical details, or indeed any details of implementation, should figure in high-level explanation. All it says is that what figures

2. Fodor’s official claim is that these properties do not appear in the laws of fundamental physics (since he takes appearance in the laws of a discipline to be necessary and sufficient for naturalness in that discipline); however, the claim that the explanatory properties of the high-level sciences do not appear in any guise in the fundamental laws seems, though undeniably true, of considerably less philosophical interest than the stronger if vaguer claim I have substituted for it in the main text.

in high-level explanations can be specified in physical terms (explanatory physicalism), and should be determined by the same principles that determine relevance in physics (explanatory monism); it is quite possible that the relevance-determining rules in physics declare some details to be irrelevant.

Second, explanatory reductionism does not entail that what is relevant in high-level explanations is also relevant in explanations in physics or vice versa: explanatory relevance is relative to what is being explained, and the events, patterns, and regularities that are explained in biology are not the same as those that are explained in physics. Apply the same rules to different explananda, and you may get very different results.³

Yet the doctrine of autonomy provides a powerful argument against explanatory reductionism, I think. It is not a deductive argument, but an inference to the best explanation: that the taxonomy of explanatory properties in physics is quite different from the taxonomies of various other sciences, and indeed that the taxonomies apparently cross-classify one another, is best explained by the denial of explanatory reductionism. If explanatory reductionism were true, the argument goes, we would feel both tempted and obliged to align the explanatory styles of, and in particular the taxonomies of explanatory relevance of, the various sciences. But we feel this pull little if at all—as Fodor observes in the quotation with which this section begins—so it is likely that explanatory reductionism is false.⁴

A reductionist could of course reject the autonomy thesis, but this is not a strategy I would recommend; the fact of autonomy seems clear enough to me.

3. Explanatory autonomy is, however, inconsistent with “reductivism” as it is defined in Fodor’s paper; for Fodor, reductivism is the thesis that every property that plays an explanatory role in the high-level sciences also plays an explanatory role in physics (indeed, that every property that appears in high-level laws also appears in the laws of physics)—but this is surely a view that no sane person has advocated in print.

4. The other major argument against explanatory reductionism—more particularly, against explanatory physicalism, its first component—is built on a direct case for the undefinability of explanatory high-level predicates, such as Kitcher’s (1984) case for the undefinability of *gene*; a defense against such arguments is given by Strevens (forthcoming).

I will rather aim to defuse the argument against explanatory reductionism by showing that a reductionist account of explanation can make sense of autonomy just as well as any anti-reductionist account.

2. The Corridor and the Campus

Before I try to explain autonomy in physicalist terms, let me make my job harder by adding some more content to the fact of autonomy. Or perhaps this is not so much an addition, as a making explicit of something that is already present in Fodor's observations and others' arguments for the autonomy thesis, namely, that the range of explanations in the sciences, and equally the range of explanatory taxonomies—the categories of properties treated as explanatorily relevant by the various sciences—exhibits many discontinuities.

Let me explain what I have in mind by providing a contrast, that is, by describing what it would be like if there were no such discontinuities. Imagine a university laid out along a single, extremely long corridor. At one end of the corridor are the researchers working on fundamental physics. At the other end are researchers working on economics or perhaps sociology. In between the two are the offices of those concerned with the intermediate levels: next to the physicists the chemists, then the molecular biologists, the cytologists, the physiologists, the ethologists, the evolutionary biologists, and so on, through the psychologists and on to the social scientists. (The one-dimensionality of the corridor is not what matters here; a vast room full of cubicles would serve just as well.)

Suppose that you make a journey from the physical to the social end of the corridor, knocking on each office door along the way and asking for a succinct summation of the occupants' explanatory practices: what details they consider to be explanatorily irrelevant, what properties or laws they consider to be relevant, the mathematical forms (if any) of their explanations, and so on.

In a universe unlike our own, such a journey might reveal the following sort of continuity in explanatory practices: as you go from door to door, the practices within change gradually and continuously, each close to, but not identical to, its neighbors. Moving one office along, you would find that a few more details were considered irrelevant, a few more variations in implementation negligible, while certain other properties go—gradually, one door at a time—from explanatorily insignificant, to marginal, to occasionally useful, to central, to essential.

In such a world, you might still have explanatory autonomy in the sense defined in the previous section: the physicists would care about details that the evolutionary biologists considered utterly irrelevant, while the biologists would cite properties, the realizers of which have nothing in common that would interest a physicist. But there would be a continuity to these differences: they would emerge slowly as the corridor was traversed.

In our own world, things are quite different. There is no transition zone between the disciplines, or between the sub-disciplines, or at least such zones are rare and sparsely inhabited. Will things change as more science is done? Will a more sophisticated neuroscience, for example, fill the gap between psychology and physiology? Perhaps, but then again, high-level scientists' lack of interest in the transition zones seems just as often to be based on principle—the zones have nothing explanatory to offer—as based on the parlous state of the science. I will not try to argue for these points here; others have done a better job than I can elsewhere. Let me simply take the university's explanatory discontinuities as given. The internal explanatory logic of the university, then, calls for its organization not as a corridor but as a campus: a collection of distinct buildings, each with its own distinctive way of going about the explanatory enterprise, and in particular, its own set of criteria for explanatory relevance, qualitatively different from those of its neighbors.

This campus structure, this picture of explanatory practice in science as a

mosaic of autonomous sub-disciplinary explanatory styles, further strengthens the argument against explanatory reductionism, pointing rather to some kind of pluralism of explanatory materials and practice.

3. Accounting for Explanatory Autonomy

A reductionist account of explanation—an account that provides a recipe for constructing the explanation of anything out of purely physical resources—is, I will show, equally able to explain autonomy as any pluralist or anti-reductionist account.

The key notion is that of explanatory abstraction, a process in which parts of a model that are explanatorily irrelevant to the explanandum are removed. Different explananda in different disciplines call for differing amounts of detail to be deleted. Further, there is a considerable degree of systematicity to the removals, resulting in distinct, discontinuous explanatory taxonomies for the disciplines.

What follows is based on my own reductionist account of explanation, which I call the *kairetic account* (Strevens 2009); however, I have formulated the philosophical explanation in such a way, I hope, that it is clear that many details of that account are irrelevant to my account of autonomy.

3.1 *The Raw Material of Explanation*

The raw material of scientific explanation, according to the kairetic account, is a fundamental-level relation of causal influence.⁵ Newtonian force is the paradigm of such an influence relation. In the Newtonian world-view, force is responsible for all changes of velocity of all objects, microscopic or massive, and so along with inertial motion, for everything that happens. It is described

5. There is another kind of material as well, something that I call entanglement, that will not play a role in this paper (Strevens 2009, §7.3).

completely by fundamental physics. The question whether there is causal influence to be found in modern physics is more difficult, but I ask you to put it entirely to one side, and to suppose that physics does indeed give us such a relation, as many philosophers of causation have argued (Reichenbach 1956; Salmon 1984; Dowe 2000; Lewis 2000).

Let me focus, in what follows, on the explanation of singular events or states of affairs. The kairetic and other similar causal theories of explanation hold that an event is to be explained by showing how it was produced by the aggregate causal influence of other states of affairs and the laws in virtue of which they exerted their influence. In a Newtonian world, for example, an event is explained by exhibiting whatever objects, events, or states of affairs pushed around certain constituents of the world in a way that realized the occurrence of the event, along with the laws in virtue of which the pushing around occurred, that is, the laws in virtue of which the relevant forces were brought to bear.

Is that all that needs to be said about explanation in science? Some writers have thought so. To explain a broken window, for example, the breaking should be understood as a certain kind of movement of the many fundamental-level particles that constitute the window, and the explanation should consist (in principle) of a complete list of the forces that made the particles move in the way that they did, along with a list of the things that exerted the forces and the laws in virtue of which they did so. This is more or less the view proposed by, for example, Salmon (1984).

Later Salmon reconsidered his view (Salmon 1997). Suppose, he imagined, that the window is broken by a baseball. Among the causal influences on the movements of the window's particles at the time of breaking are both the impact of the ball on the glass and the shouts of the ball-players (which cause the pane to vibrate). But the shouts, Salmon reasoned, are explanatorily irrelevant: it is the impact of the ball alone that explains the breaking.

Not all causal influences on an event are explanatorily relevant to the

occurrence of that event, then. What additional criterion for relevance distinguishes the explainers from the rest? The later Salmon, the kairetic account, and a number of other approaches to explanation, favor the idea that the ball's impact is relevant to the breaking, while the players' voices are not, because the impact but not the voices *made a difference* to the fact that the window broke. While the voices affected the movement of the window's particles, that is, they did not affect them in the sort of way that tipped the balance between the window's breaking and its not breaking.

How to understand balance-tipping, or difference-making? There are several possibilities. Salmon suggested that it is a matter of statistical relevance: the impact makes a difference because it, but not the players' voices, raised the probability of the breaking. Other writers have suggested a counterfactual test: the impact but not the noise makes a difference because, had there been no impact (just noise), there would have been no breaking, whereas had there been no noise (just the impact), the breaking would have occurred all the same (Woodward 2003).

The kairetic account puts forward a third possibility, loosely related to Mackie's test for causation: removing the impact but not the noise from the model of the breaking process makes it no longer a model for breaking, while removing the noise but not the impact has no such effect.⁶ More particularly, an explanatory model for an event may be constructed as follows. Start with a complete enumeration of the causal influences that account for the motions making up the event. This enumeration will constitute a model for the occurrence of the event in part because it entails that the event occurred (assuming determinism, as without any further comment I will). Now remove as much information as you can from the model without invalidating the entailment of the event's occurrence. Such removal will consist either in deleting all reference to a particular causal influence, leaving open the question whether the influence was present or not, or more often, in replac-

6. Reasons to favor this approach over the others are given in Strevens (2009, chap. 2).

ing a detailed description of a number of influences with a more abstract description—as when, for example, you replace a precise specification of a baseball’s velocity and mass with the statement that these parameters fell within a certain range. When you are done—when you have removed all the detail that you can—you will have a model for the event that represents only difference-making features of the world, thus that contains only explanatorily relevant information.

There is much more to be said about this process of abstraction by which the kairetic account refines a mass of causal detail into relatively abstract explanatory model; I refer you to Strevens (2009, chap. 3) for the details. One thing I will say here: the account has much the same sort of thing to say about the explanation of regularities and laws that it has to say about events. Regularities and laws are to be explained by exhibiting causal mechanisms; all such mechanisms should be pruned in the same way as the causal models that explain events, so that they are as abstract as they can possibly be while still entailing that the explanandum holds, that is, that the regularity obtains.

3.2 *Objective Irrelevance*

Suppose that you have a causal account of explanation based on the notion of difference-making, whether difference-making is understood in terms of statistical relevance, counterfactual dependence, or the kairetic account. Such an account will discriminate between events that make a difference to another event and those that do not—saying, for example, that the impact of the baseball made a difference to the window’s breaking but the players’ shouts did not. It will also discriminate between features of a single event that make a difference and those that do not. Though the fact that baseball hit the window hard made a difference to the breaking, the fact that it had a cork rather a rubber center made no difference. (Salmon: the probability of breaking is the same given a rubber or a cork center. Woodward: if the cork center had been switched out for a rubber center immediately before

impact, the window would still have broken. Strevens: an explanatory model that abstracts away from the ball's material—that specifies only that the ball's center had a certain rigidity and mass—will still entail that the breaking occurred.)

To put it another way, according to any of these views there are certain details of implementation that make no difference to the breaking. They have a causal influence on the breaking, but this influence makes no difference to whether or not the breaking occurs, only to exactly how it occurs (perhaps the shards of the window would have sprayed out somewhat differently had the ball been rubber- rather than cork-centered). These details of implementation are explanatorily relevant, then, and should, like the players' voices, be omitted from an explanation of the breaking.

Observe that explanatory relevance is relative to the explanandum. If you are explaining why the window broke, then the details of the “implementation” of the ball—its core material—are irrelevant. However, if you are explaining some more particular fact about the breaking, such as the fact that the broken glass was found such and such a distance from the window frame, then the same details may well be relevant.

In general, although not exceptionlessly, the more finely grained the explanandum, the greater the amount of finer-grained detail that is relevant to its explanation. This goes a long way toward accounting for the evident indifference shown by the high-level explanatory enterprises toward details of physical implementation:

1. The exact positions of the rabbits in an ecosystem makes no difference to the rate of predation in the system; therefore, population ecologists need not attend to precise rabbit locales.⁷
2. The exact positions of the many molecules involved in the process of meiosis make no difference to the process's conforming to Mendel's sec-

7. This observation is qualified and defended by Strevens (2003).

ond law (properly qualified); therefore, a cytologist explaining the law may ignore, for example, the differences between individual telomeres.

3. The positions of small asteroids, interplanetary debris, and dust make no difference to the planets' obeying Kepler's laws; therefore, explainers of the laws need not attend to anything much smaller than a planet. Further, the exact positions of the molecules in a planet make no difference to its obeying the laws; therefore, the explainers need not attend to the details of planetary construction.

This is not to say that the details in question do not make a difference to *something*: for every detail, no matter how small, there is an explanandum of such exquisite fineness of grain that it has among its explainers, its difference-makers, that detail. What the details do not make a difference to are a wide range of relatively high-level phenomena.

Furthermore, there is a systematicity to these facts about difference-making and irrelevance. It is not that dust makes a difference to the orbits of some planets but not to others, or that the position of rabbits makes a difference to the rate of predation in some ecosystems but not others, or for that matter that the position of rabbits makes no difference to rabbit predation rates but the position of pigeons does make a difference to pigeon predation rates: the non-difference-making, the irrelevance, is across the board. High-level explainers may consequently ignore certain kinds of detail as a matter of general policy; they need not treat each case on a custom basis.

Approaching the question of orbits, for example, the celestial mechanic has good reason in advance to abstract away from planetary construction and the existence of interplanetary rubble and dust. For explanatory purposes, they need not see their system as made up of molecules at all; it is enough to see it as made up of planets—discrete spherical objects of great mass and indeterminate composition. The facts about difference-making, then, provide the explainer, by way of abstraction, with a high-level taxonomy of

the system in question from which certain aspects of the fundamental physics have disappeared altogether.

Likewise, the cytologist is apt to think in terms of telomeres rather than in terms of DNA sequences, let alone in terms of the individual molecules that make up such sequences, each with its own particular position, velocity, modes of vibration and so on, and the population ecologist will naturally think in terms of populations or standard types rather than actual organisms in all their individuality, let alone in terms of the organisms' molecular makeup.

The systematicity of difference-making, in short, not only requires a high-level explainer to ignore many matters of physical implementation; it encourages them to approach their explanatory endeavors with a distinct style, a taxonomy of entities and a set of relations among those entities whose level of abstraction is precisely tailored to fit the characteristic explananda of the high-level domain in question. Hence a compartmentalization of the explanatory enterprise—hence the campus, not the corridor.

3.3 *Contextual Irrelevance*

If a detail makes no difference to an explanandum, say that it is *objectively* irrelevant to the explanation of that explanandum. It is compulsory, when explaining, to ignore objectively irrelevant factors. There is another reason for ignoring details that is not compulsory in quite the same way, but that is equally important in accounting for explanatory autonomy; I call it *contextual* irrelevance.

Suppose that alien anthropologists of technology descend on planet earth and attempt to reverse-engineer the internal combustion engine—to construct an explanation of how the thing works. They might reasonably pursue their task by dividing the engine into (apparent) functional units: the piston assembly, the starter motor, the cooling system, and so on. A team would be assigned to each unit with the task of explaining how that unit contributed to

the functioning of the whole. The advantage of such a division of labor is, of course, that one part of the engine can be analyzed without any detailed understanding of the other parts. The other parts cannot be ignored altogether, but they can be treated as “black boxes” with specified inputs and outputs, but no internal details. The team working on the piston assembly needs to know that there is a cooling system, and something about the parameters of that system’s operation—how it reacts to increasing engine temperature, how fast it cools, and so on—but that is all.

The full explanation of the engine’s workings is attained by bringing together the teams and replacing each black box with the proper internal causal model of its workings (omitting objective non-difference-makers). Until that time, no individual fully understands the engine. This is a good thing: it would be a mistake, in a practical sense, for the piston people to start thinking about cooling. The rules of the “division of cognitive labor game” might therefore be stated as follows: every group has an explanatory task, to analyze a particular functional component of the engine; the details of any other component’s workings are to be considered irrelevant to the explanatory task.

This is what I mean by contextual irrelevance in explanation. The workings of the cooling system are objectively relevant to explaining the engine’s behavior, but they are in another sense irrelevant within certain contexts of inquiry, such as the context in which the investigation of the piston assembly takes place. Contextual relevance is very different from objective relevance. Whereas the facts about objective relevance are decided by the world, the facts about contextual relevance are decided by us, the organizers of inquiry into the world. Whereas including objectively irrelevant factors in an explanation is an intellectual error, including contextually irrelevant factors is a social or practical error. Nevertheless, the success of an explanatory investigation may hinge just as much on the norms of contextual relevance as on the norms of objective relevance. Investigators should take both equally seriously. (I

should perhaps add that the norms for both kinds of relevance govern what goes into an explanation, not what the explanatory investigators may think about. A successful explainer will of necessity have to spend rather a lot of time thinking about non-difference-makers, just so as to be able to recognize them as such.)

The phenomenon of contextual irrelevance in explanation manifests itself in many endeavors besides functional analysis. When I ask “What explains the fire?” and you reply “Among other things, the presence of oxygen”, you have provided a non-explanation, not because you have violated a norm of objective relevance—the oxygen most certainly is a difference-maker for the fire—but because you have violated a norm of contextual relevance. We both know that oxygen was present and made a difference; the explanatory task I set you in my question is to provide me with other parts of the causal story.

More important to this paper, science divides its cognitive labor in much the same way as the alien anthropologists. The division appears in functional analysis, as in the study of the physiology of the human body, where the immune-system specialists black-box (for example) the parts of the brain controlling the respiratory system. But even more salient is the individuation of the explanatory pursuit by levels: the physicists study the fundamental particles, the chemists the molecules from which those particles are made, the cytologists the cells from which the molecules are made, and so on. Typically, investigation at one level black-boxes what is going on at lower levels— not ignoring lower-level goings-on altogether, but introducing them by way of functional specifications, as when chemists (though not quantum chemists) talk in terms of bonds or nuclei, without asking how a chemical bond is implemented or a nucleus put together, or population ecologists talk in terms of predators and prey, without trying to represent the mechanics of predation.

One motivation for black-boxing is objective irrelevance: if the details of a lower-level process do not make a difference to a high-level explanandum, they should be omitted from the explanation. But not all black-boxing of the

lower level can be accounted for in this way, I think. Consider, for example, a population ecologist's explanation why a boom in the predator population can lead, down the line, to a boom in prey population: the predator boom causes a crash in both populations, but the prey recover more quickly. The ecologist may make use of an explanatory model that represents the rate of predation as increasing with the number of predators, but which supplies no further information about the way in which predators hunt for prey. Such a model black-boxes predation: it specifies the effect of predator numbers on prey numbers, but says nothing about the mechanism responsible for the effect. Yet the mechanism is not irrelevant to the explanation in the objective sense: if you are trying to understand the relation between population booms and busts, and it turns out that the relation between predator population and predation rate plays an important role, it is far from irrelevant to ask: "Why is the rate of predation proportional to the number of predators?" If the predation mechanism does not appear in the explanation, then, it must be because the population ecologists have been excused, by some principle of division of labor, from having to supply it.

Let me turn this into a general proposition: the principle by which the economists black-box psychology, the psychologists black-box neuroanatomy, the evolutionary biologists and population ecologists black-box various aspects of physiology, and so on, is the principle of contextual irrelevance. It represents the division of the explanatory enterprise into many different parts, in this case layers, for the sake of more efficiently producing complete explanations, that is, explanations in which all objective difference-makers, whether described at the level of physics, cytology, psychology, or whatever, are brought together into a single explanatory model. (Or notionally brought together—since the sciences are all works in progress, the grand union is typically more of an aspiration than an actual accomplishment.)

It is one of the great glories of modern science that we have something approaching this full understanding in many cases, in the sense that in many

cases we can at least sketch answers to the chain of explanatory questions leading all the way down to the fundamental level. But of course, no one person is custodian of this understanding. It is spread across the campus, with (in the rabbit case) the population ecologists leading off, the cognitive ethologists then taking up the baton, and so on all the way to the particle physicists. It would be madness to organize explanatory inquiry in any other way. Science calls for the division of labor, for specialization, just as urgently as the manufacture of pins.

Objective explanatory relevance is, I proposed above, systematic; this systematicity accounts in part for the campus-like structure of the explanatory enterprise. Contextual explanatory relevance is also systematic: there are long-lasting, across-the-board rules determining what is and is not contextually relevant within an explanatory domain.

This may sound dubious. What is contextually relevant is determined by the way in which the explanatory enterprise is parceled out among the scientific community. These allocations are a matter of epistemic policy; why expect such policies to be uniform over time, from place to place, or across a discipline? If policies were simply a matter of fashion, systematicity would be surprising. But given their goal of investigative efficiency, they are strongly constrained by the world out there: there are certain divisions of explanatory labor that are very efficient, and many that are grossly inefficient.

Consider again, for example, the investigation of the internal combustion engine. One way to divide the work of understanding the engine is for one team to take the left half, one team the right half. The left-hand team might find themselves, then, trying to understand the workings of the left-hand side of a piston while black-boxing its right-hand side. This is possible in principle, but the black box in question, which must specify all behavior of the black-boxed system that makes a difference to the behavior of the system under investigation, will have to contain an extremely detailed description of the right-hand side—so detailed, that it will hardly be a black box at all.

More generally, black-boxing is useful only when the system in question is somewhat modular, in the sense that it can be divided into units, relatively few of the goings-on in one of which make a difference to what goes on in the other. (They may make a big difference; what is important is that a specification of the difference-making behavior be compact and tractable.) When the explanatory task is divided into sub-tasks corresponding to the modules, black-boxing then allows the researchers working on each sub-task to proceed knowing relatively little about the other researchers' discoveries, because they have in the short black-box specifications everything they need to know. The facts about difference-making, which determine whether a putative black box can be specified compactly or not, therefore determine which decrees about contextual relevance can profitably be made.^{8,9}

Here is what we have so far: contextual irrelevance depends on the efficient allocation of cognitive labor, which depends on opportunities for compact black-boxing on offer, which depend in turn on explanatory difference-making. But the facts about difference-making are, as we saw above, systematic.

It follows that if there is any contextual relevance at all—if there is any division of explanatory labor at all—it will tend to be systematic itself. The

8. As always with matters determined by difference-making, the length of the necessary functional specification is relative to the explanandum. If you are trying to explain why the engine weighs what it does, having one team determine the weight of one half and the other team determine the weight of the other will not create pointlessly complex black boxes, since for the purpose of understanding one half's weight nothing at all need be known about the behavior of the other half—one half's weight makes no difference to what the other half weighs.

9. The sense of difference-making relevant to determining the proper level of detail in a black box's functional specification is not quite identical to the objective difference-making discussed in the previous section. The core idea is the same: an element in a functional specification in a black-boxing explanatory model makes no difference to the explanandum if it can be removed from the model without invalidating the model's entailment of the explanandum. The difference (which is of strictly local interest) is that this kind of difference-making is not transitive, in the sense that nothing about the black-boxed system counts as a difference-maker for the specification.

same details of physical implementation will be ignored for a wide range of explanatory tasks in a high-level domain; consequently, the domain will acquire a certain explanatory taxonomy and a certain distinctive explanatory style. The university will consist not of an explanatory cultural continuum, but of a mosaic of explanatory sub-cultures—it will be a campus, not a corridor.

4. Concluding Economic Postscript

Why is there explanatory autonomy? Why do the higher level sciences seem so unconcerned with details of physical implementation; why do they have their own explanatory taxonomies and styles of explanatory narrative? Why are there discontinuities rather than smooth transitions from the characteristic explanatory style for one kind of explanandum to the style for another?

Perhaps each high-level domain has its own explanatory material—its own kind of causal and other relations—or its own rules for the adequacy of explanatory models, or both. Perhaps explanatory reductionism is false.

But then again, even if explanatory reduction is true, and in particular, if the kairetic account or something like it is correct, you would anyway get explanatory autonomy in abundance. First, you would find that many details of lower-level implementation are considered irrelevant to high-level investigations, some because they are objectively irrelevant—they make no difference to the explanandum's obtaining—and some because they are contextually irrelevant, meaning that they have been farmed out for reasons of efficiency to a complementary investigation. Second, you would find that, because of the systematicity of difference-making, which directly affects objective relevance and indirectly affects contextual relevance by determining where black box boundaries may effectively be drawn, the same details of implementation are irrelevant, whether objectively or contextually, across entire high-level explanatory enterprises or sub-disciplines. Because the matter of which details

are irrelevant depends on difference-making, and difference-making depends on the kind of explanandum in question, different sub-disciplines will ignore different kinds of details. Each sub-discipline, then, will have its own explanatory taxonomy—the set of objects and properties existing at the level of abstraction where all characteristic difference-makers and no characteristic non-difference-makers have been removed—and its own explanatory style, or way of using its proprietary taxonomy to construct explanatory models.

Let me conclude with an example discussed by Fodor, that of currency. As Fodor points out in a quotation at the beginning of this paper, economists have very little interest in the physical properties of currency. Why should this be?

First, many aspects of the physical implementation of currency are objectively irrelevant to the usual economic explananda—the color of a banknote’s ink, its size, and so on. Second, even those few physical aspects that matter, such as a banknote’s having a certain physical resilience or its being difficult to forge, are black-boxed by economists, thus left to other experts (presumably employed mainly by the mint).¹⁰

What is left if economists abstract away from everything physical about currency? The answer, of course, is a propensity that does not belong to the currency itself but to the currency’s users, namely, their willingness to deploy the currency in certain kinds of exchange. Here, too, economists do much abstraction: they typically pay little attention to the implementation of the psychology of exchange, since the psychological details are either objectively irrelevant or are black-boxed. But something remains, in the form of the preference rankings for bundles of commodities that form the basis of modern microeconomics.¹¹ These dispositions to exchange are causal dispositions, realized in fundamental-level relations of causal influence, though almost all of

10. Funnily enough, Fodor’s favorite example of an economic law, Gresham’s law, may be the only economic regularity that does turn to some degree on the physical implementation of currency.

11. In fact, microeconomists typically abstract away from currency altogether.

the details of the realization are objectively irrelevant to economic explananda and the remainder are black-boxed—left to the psychologists to figure out (who leave something over in turn for the neuroscientists, who don't neglect to pass on some problems of their own to the molecular biologists, and so on).

The proponents of the new behavioral economics would like to convince their colleagues that a little too much is omitted in such abstractions: more psychological detail would, the argument goes, considerably improve economists' explanations. They make their case not by pointing out that economic decisions are implemented in human psychology—that is hardly news, and cuts no ice in an explanatorily autonomous discipline such as economics—but by observing that certain quirks of the psychology stand to make a difference to economic decision-making, and so are relevant to economic explananda. Explanatory reductionism can make good sense of this tactic: it is the explanatory connections between levels, as manifested in relations of difference-making, rather than the vagaries of some irreducible and separate explanatory style or fashion, that determine when a high-level science must dirty its hands with matters of implementation and when it may remain aloof.

It is possible, then, to endorse everything that Fodor says about currency and economic explanation—the lack of concern with physical implementation, economics' *sui generis* explanatory style—while holding on to the most full-blooded version of explanatory reductionism. The same is true, I suggest, for the other high-level sciences. Explanatory autonomy is to be expected regardless of whether explanatory anti-reductionism is true.

References

- Dowe, P. (2000). *Physical Causation*. Cambridge University Press, Cambridge.
- Fodor, J. A. (1974). Special sciences. *Synthese* 28:97–115.
- Kitcher, P. (1984). 1953 and all that: A tale of two sciences. *Philosophical Review* 93:335–373.
- . (2001). *Science, Truth, and Democracy*. Oxford University Press, Oxford.
- Lewis, D. (2000). Causation as influence. *Journal of Philosophy* 97:182–197.
- Reichenbach, H. (1956). *The Direction of Time*. University of California Press, Berkeley, CA.
- Salmon, W. C. (1984). *Explanation and the Causal Structure of the World*. Princeton University Press, Princeton, NJ.
- . (1997). Causality and explanation: A reply to two critiques. *Philosophy of Science* 64:461–477.
- Sober, E. (1983). Equilibrium explanation. *Philosophical Studies* 43:201–210.
- Strevens, M. (2003). *Bigger than Chaos: Understanding Complexity through Probability*. Harvard University Press, Cambridge, MA.
- . (2009). *Depth: An Account of Scientific Explanation*. Harvard University Press, Cambridge, MA.
- . (Forthcoming). The explanatory role of irreducible properties. *Noûs*.
- Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford.