# The Whole Story:
## Explanatory Autonomy and Convergent Evolution

Michael Strevens

To appear in *Explanation and Integration in Mind and Brain Science*,
edited by David M. Kaplan, Oxford University Press

Abstract

Causal explanations in the high-level sciences typically black-box the low-level details of the causal mechanisms that they invoke to account for their explananda: economists black-box psychological processes; psychologists black-box neural processes; and so on. Are these black-boxing explanatory models complete explanations of the phenomena in question, or are they just sketches of or templates for the whole explanatory story? This chapter poses a focused version of the question in the context of convergent evolution, the existence of which appears to show that underlying mechanisms are completely irrelevant to the explanation of high-level biological features, including perhaps thought and language—in which case a black-boxing model would be a complete explanation of such features rather than a mere sketch. Arguments for and against such a model's explanatory completeness are considered; the chapter comes down tentatively against.

## 1. Explanatory Disintegration

Look wherever you like in the higher-level sciences—in cognitive psychology, or economics, or anthropology, or even much of biology—and you will find explanatory models that are entirely unconcerned with lower-level mechanisms. In economics, you find models of the consequences of economic decision-making that have nothing to say about the psychology of decision; in psychology you find models of decision-making that have nothing to say about the way that psychological processes are implemented in the cerebral substructure; in neuroscience you may, depending on your corridor, find quite a bit of cytology or chemistry, but typically no quantum chromodynamics.

This absence of the lower level is one aspect of what is called the explanatory autonomy of the high-level sciences. Explanatory autonomy is perhaps itself only one kind of autonomy, to be set alongside methodological autonomy, metaphysical autonomy, managerial autonomy, and so on. I focus on explanation in this chapter because it is raises the problem of the integration or the unity of the sciences in principle and in the long term, the production of explanations being a scientific end and not merely a means.

The autonomy, or disunity, or disintegration of the scientific disciplines and sub-disciplines poses a prima facie challenge to those of us who believe that we live in a coherent world and that science's overriding task is to give us a clear picture of that world. If the subject matter is a unified whole, why is its scientific portrait so fragmentary?

Perhaps the world is not so unified—perhaps it is dappled (Cartwright 1999) or disordered (Dupré 1993). It might be, for example, that the theories of a completed cognitive psychology could not be translated into, or otherwise explanatorily related to, the language of a completed neuroscience. To try to fit the two together would then be like solving a jigsaw puzzle made up half from one set and half from another set of differently shaped, differently cut pieces. Or it might be that psychological theories can be translated into neuro-argot, but that the resulting sentences cannot be derived from existing

neuroscientific theories, either because of some sort of emergence or perhaps because the theoretically interesting categories of neuroscience cross-classify the interesting psychological categories.

If any of these visions of disunity is correct, then the present-day autonomy of the sciences of the mind would be a sign of maturity: anticipating insurmountable barriers to integration, the sciences have renounced vain pretensions to a seamless theory of thinking; each consequently pursues its own ends in its own way.

The balkanization of our representations of the world need not, however, imply a balkanized world. The sum total of being might be the integrated whole imagined by Plato or Spinoza, yet our windows onto the world might be for some reason manifold and variously shaped and tinted.

In this chapter, I juxtapose two such reasons. According to the first, the many windows exist for practical reasons, to better organize the process of uncovering the structure of a unified world. On this sort of view, explanatory autonomy is a temporary arrangement: a completed high-level science will pay just as much attention to, will be just as constrained by, and will derive at least as much of its explanatory richness from low-level structures such as underlying mechanisms as to, by, and from principles of high-level organization.

According to the second, the many windows exist because of the nature of explanation itself: the lower-level facts are irrelevant, explanatorily speaking, to the higher-level facts. On this sort of view, autonomy in present-day explanatory practice reflects the inherent structure of explanatory knowledge. The high-level sciences neglect low-level mechanisms for principled reasons, and will continue to do so even in their finished form. They need not, and indeed should not, draw on the lower-level sciences for their explanatory content, nor need they be constrained by the lower-level sciences' explanatory organization of things.

At the heart of the chapter is an argument, based on convergent evolution,

to prefer the second picture to the first. I do not endorse the argument; rather, my aim is to develop it and to investigate possible responses on behalf of those thinkers who feel the explanatory pull of underlying mechanisms.

## 2. One World; Many Sciences

Autonomous explanatory practices in a unified world: why? Let me describe the two possible answers to be investigated in this chapter in more detail.

According to the first answer, the relative lack of integration between the higher and the lower-level sciences is motivated by the practical benefits of intellectual specialization.

Suppose, for example, that in order to produce a complete explanation of some economic phenomenon, we need both an economic story relating the explanandum to various patterns of decision-making and a psychological story that accounts for those patterns (by relating them to more basic principles of thought, which are to be accounted for in turn by the topology of neural interchanges, their physical substrate, and ultimately fundamental-level physics). The economic and the psychological explanations are, in that case, two panels of a larger mural—the big picture that constitutes the whole story, the full explanation of why that phenomenon occurs.

Each part of the mural draws on particular skills. Some require advanced mathematics, some the manipulation of temperamental experimental setups. Better, then, to divide the work among teams each specializing in intellectual labor of that variety—to give the game theory to the economists, the laws of mental logic to the psychologists, the brain circuitry to the neuroscientists, and so on.

The economists are ignoring the psychologists, in that case, not because psychology is irrelevant to their explanatory enterprise, but because the efficient division of labor requires a certain intellectual compartmentalization: the point is precisely that by not thinking about the psychology, however relevant it may be to the economic master narrative, you make yourself a

better because more focused game theorist, and so contribute to the narrative a more perfect game-theoretic tableau. It is only one thread among many in the explanatory action, but by neglecting the thread's final purpose, you make a stronger, more flexible, more colorful contribution to the tapestry of knowledge. By your deliberate neglect of the other strands of the big story, you contribute more surely, more quickly, more reliably to its telling.

On this view of things, once the explanatory work is done, there is notionally a great gathering. Each research group, each department, comes bearing its own particular pieces of explanation, and then as the assembled scientists watch, the parts are assembled into an explanatory entirety. Contemplating the big picture, the whole story, each researcher at last, for the first time, understands fully the phenomena they have been studying all their life.

What are these "pieces of explanation"? They are what Hempel (1965, §4.2) called partial explanations or explanation sketches, that is, legitimate explanations from which some pieces or details are omitted.[1] The omission takes a particular form in the explanatory products of the high level sciences: descriptions of mechanisms are replaced by black boxes, that is, very roughly, job specifications paired with assertions that something or other gets the job done. The game-theoretic economist specifies, for example, that something in people's heads computes the optimal course of action in such and such a context, and something else makes the plan a reality, without saying what these somethings, these underlying mechanisms, are or how they work.

Because a model of the underlying mechanisms is nevertheless necessary for full understanding of the economic phenomena, practically inspired black-boxing results in an explanation that is at best partial, at best a sketch. It leaves

1. A *partial explanation* in Hempel's sense is a complete explanation of some aspect, but not every aspect, of the explanandum. A complete explanation of why Mount Vesuvius erupted in 79 CE is therefore a partial explanation of why it erupted in October of that year; it would become a complete explanation of the October eruption were sufficient details added to the explanatory model to entail the October date. An *explanation sketch*, by contrast, is not a legitimate explanation of anything, but rather an outline or template or even just a fragment of a complete explanation.

an epistemic hole, but is desirable all the same so that science may enjoy the efficiencies made possible by the division of labor.

There are two components to the view I am here describing, then: first, the thesis that explanations are incomplete without models of the relevant underlying mechanisms, and second, the thesis that although black-boxing explanations are explanatorily unfinished, the most efficient organization of explanatory inquiry will mandate the production, for purely practical reasons, of precisely such things—with the rider that there is a further phase in scientific explanation construction in which the various explanatory parts are woven together to create the whole explanatory story.

Who advocates such a view? That a description of underlying mechanisms completes an explanation, or to put it another way, that describing such mechanisms increases a model's explanatory value, is a postulate popular among those who hope to integrate psychology and neuroscience. Piccinini and Craver (2011, §7) write that in the explanatory enterprise "full-blown mechanistic models are to be preferred"; Kaplan (2011, §2) also favors explanations that "delineate the underlying mechanisms". For these writers, cognitive psychologists produce explanatory templates that are to be filled in, when the time comes, by a mature neuroscience.

I myself think along the same lines. An explanation that black-boxes is leaving out something explanatorily important; at the same time, black-boxing is for reasons of efficiency ubiquitous in the high-level sciences—not just in the cognitive sciences but everywhere, down to and including much of physics (Strevens 2008, §5.4; Strevens 2016).

If this is correct, then the fact of explanatory autonomy—the fact that explanatory inquiry in the sciences is modular, even fragmented—provides no more reason to infer that the world itself is fragmented, than the modularity of the various parts of a Shenzhen assembly line provides reason to think that there is no finished product.

* * *

There is another way to reconcile explanatory autonomy with a unified world that makes no appeal to practical considerations; it will serve as the chief rival to the practical view in this chapter.

On this second view, the lower-level details ignored by a high-level inquiry are typically explanatorily irrelevant to the phenomena under investigation. An economist's neglect of psychological details, for example, is on this approach due to the irrelevance of the mind's decision-making mechanisms to economic phenomena. That is not to say that nothing about the mind is relevant to economics, but rather to say that what is relevant is captured by the appropriate black box: it matters *that* the mind finds the optimal move in the game, but not *how* it finds that move. The explanatorily best economic model will therefore contain a black box asserting the *that* without describing the *how*.

One level down, the story is repeated: the black-boxing of brain-related details by a cognitive psychologist is, far from an embarrassment, an omission mandated by the canons of explanatory relevance. In explaining some cognitive capacity, it is highly relevant that the thinker is using this inferential rule rather than that rule, but irrelevant how the all-important rule is implemented. Thus, even a completed cognitive psychology will float atop the seething neural sea.

According to this view, then, the high-level sciences are explanatorily autonomous from one another and from the lower-level sciences because they attempt to explain different sets of phenomena and because the standards of explanatory relevance judge each of these sets of phenomena irrelevant—except in black-box form—to the explanation of most or all of the others. Among philosophers who think this way are Franklin-Hall (forthcoming) and, to some extent, Garfinkel (1981). I myself believe that it has something to offer, though it falls far short of accounting for all instances of black-boxing in the high-level sciences, since most are simply a matter of labor division.

A related view attributes autonomy not to a single standard of explanatory

relevance making different judgments about different classes of explananda, but to different domains having distinct standards for relevance. Thus, there is not a single relevance-determining principle that rules the details of psychological mechanisms irrelevant to economic phenomena and the details of neural mechanisms irrelevant to psychological phenomena. Rather, the economists have their own, distinctive, idiosyncratic relevance principle that discriminates against psychology, while the psychologists have a rule, different from the economists', that discriminates against neurons in turn. For dialectical purposes in what follows, the two relevance approaches—one positing a single standard for relevance and one positing a standard for every scientific domain—can be lumped together. I take the simpler single-standard version as my paradigm.

* * *

How to distinguish these two explanations of autonomy? Is autonomy a manifestation of the division of cognitive labor, or is it legislated by the canons of explanatory relevance? Or neither?

Many well-known examples of explanatory autonomy seem to be accounted for equally well on either view. Economists, as Fodor (1974) remarked, show a studied neglect of the finer details of the mechanisms of currency circulation. They have nothing to say, for example, about the machinery used by automated tellers to dispense banknotes, or about the queuing system inside the bank, though both may play an important role in bank runs. While this certainly establishes that the high-level sciences are uninterested in calling the plays molecule by molecule, it does not reveal the foundation of their disregard. Are they simply leaving the details to the paper engineers and retail consultants so that they themselves can focus more intently on the workings of their macroeconomic models, though they recognize that both kinds of facts are part of the complete explanation of the near-collapse of, say, the Northern Rock bank? Or do they think that the details are irrelevant, that anything beyond some simple black boxes would add nothing of explanatory

value, would in no way enhance our understanding, of those economic events of 2007?

The fact that the details are plainly ignored in economics departments is easily accommodated on either approach, since even if the details of the queuing system were explanatorily relevant, it would make very good practical sense to organize the study of macroeconomic models separately from the study of customer service in retail banking. The truth about autonomy cannot be read off the surface features of scientific practice.

How to make progress, then? A philosopher like me will consult their intuitive judgments about explanatory relevance. Does it seem that the queuing system is explanatorily relevant? That the engineering of the ATMs is relevant? Or to take some more serious cases, does knowledge of the flow of neurotransmitters add to my understanding of adult human causal reasoning? Does it add to my understanding of the connection between printing money and inflation?

The answers are a mixed bag. Many low-level details about the implementation of high-level processes seem clearly to be irrelevant to the explanation of phenomena brought about by those processes, because as explained below, they make no difference to the phenomena's occurrence. But for some low-level details, matters are not so clear. It would be useful to have an argument pointing one way or the other that did not hinge on intuitions about relevance. In the next section I present an argument for the complete irrelevance of the low level that offers as evidence not intuitions, or even scientific practice, but the very structure of the living world.

## 3. Convergent Evolution and the Irrelevance of Mechanism

Golden moles, which comprise the twenty-one species of the family *Chrysochloridae*, are small southern African animals that live an almost entirely subterranean life. They are marvelously well adapted to existence underground, with their tightly packed fur that slides through sand and soil keeping dirt and

water at bay and their short and powerful legs tipped with claws apparently tailor-made for excavation. They have eyes that do not see, and tiny earholes that barely hear; touch is their sensory guide to the world.

Marsupial moles, which comprise the two species of the genus *Notoryctes*, live a life not unlike that of the golden moles, and it shows. They have many of the same adaptations—the fur, the claws, the lack of sight—and perhaps most striking of all, their overall aspect is remarkably similar to the golden moles (figure 1).

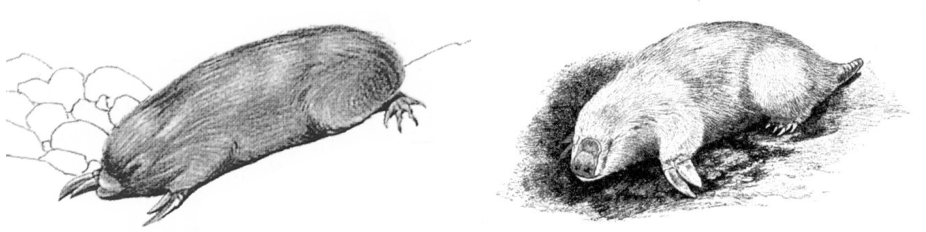

*Figure 1:* Golden mole (left); marsupial mole (right). Golden mole drawn by Clare Abbott, from P. Apps, *Smithers' Mammals of Southern Africa: A Field Guide*, Struik Nature, 2012. Reproduced by kind permission of Penguin Random House South Africa. Marsupial mole from F. Beddard, *The Cambridge Natural History Volume X: Mammalia*, Macmillan, 1902.

The two taxa are not at all evolutionarily related, however, or at least, they are less related than any two placental mammals. (Nor are the golden moles at all closely related to the true moles, which make up most of the family *Talpidae*—they are, rather, relatives of the tenrec. I will nevertheless continue to refer to golden moles and marsupial moles as "moles", thereby using the term morphologically rather than phylogenetically.) Their similarities are due to convergent evolution: faced with similar environmental challenges, natural selection has fabricated similar phenotypes.

There are multitudinous other examples of convergent evolution. Some of the most conspicuous are cases in which there are marsupial and placental versions of the same mammalian body plan: the thylacine (Tasmanian tiger) and the placental canids (wolves, jackals, and so on); the kangaroo and the

Patagonian mara; the marsupial mulgara and the mouse; the marsupial sugar glider and the placental flying squirrel. There are monotreme, marsupial, and placental anteaters all of which have evolved claws for tearing open anthills or termite mounds and long sticky tongues for scooping up their swarming inhabitants. I could continue with examples of convergent evolution in other vertebrates, other phyla, other kingdoms, or at the molecular level—but let's move ahead with the moles.

* * *

In the converging contours of the moles, nature itself seems to have written the answers to the questions about what matters and what does not matter to the molding of biological form, telling us what is relevant and what is not in explaining phenotypic structure. The low-level biological cogs and levers are evidently of very little importance in deciding the overt physiology of the moles. Placental or marsupial, true mole or tenrec fellow traveler—it is all, from selection's perspective, the same. Provided that the physiological substrate satisfies a few broad conditions that might easily be represented by a black box, the adaptive advantage of the phenotype for underground living is sufficient, acting alone, to make all moles alike. Let me try to capture this intuitive sense of the irrelevance of the causal underlayer in the form of a philosophical argument.

Call the various properties shared by the golden and the marsupial moles the *talpid phenotype*.[2] Call the mode of living shared by the two kinds of mole, made possible by the shared features of their environments, the *fossorial lifestyle*. I want to run the following argument on behalf of the explanatory account of autonomy, that is, the view that the high-level sciences for the most part ignore underlying mechanisms because they are objectively explanatorily irrelevant.

2. *Talpa* is the Latin for mole; as noted above, the family centered around the true moles is the *Talpidae*.

1. The complete evolutionary explanation of the talpid phenotype in both golden moles and marsupial moles is the same, adverting to the adaptive advantages of the phenotype for creatures living the fossorial life.

2. The underlying mechanisms involved in the evolution of the golden moles' talpid phenotype are in many cases different from the underlying mechanisms involved in the evolution of the talpid phenotype of the marsupial moles.

∴ The complete evolutionary explanation of the talpid phenotype in either group excludes the details of all underlying mechanisms—or else the complete explanations, differing with respect to these mechanisms according to premise (2), would be non-identical, contradicting premise (1).

If the underlying mechanisms do not appear in the complete explanations of the talpid phenotypes in golden and marsupial moles, then they are explanatorily irrelevant. An explanation of the phenotype that described some aspect of the mechanisms would be making a certain sort of explanatory error.

The argument does not entail that every underlying mechanism is irrelevant to the explanation of the talpid phenotype; the mechanisms that are shown to be irrelevant are only those not shared by golden and marsupial moles. But there is no obvious reason to think that the shared underlying mechanisms are any more relevant in principle than the rest, so the argument suggests, without implying, that the complete explanation of the talpid phenotype in golden and marsupial moles is comprehensively black-boxing: it contains not a single underlying mechanism. That itself provides a powerful reason, in the shape of a formidable paradigm, to think that the high-level sciences' principled disregard of low-level mechanisms is attributable to explanatory, and not merely to practical, reasons.

Kitcher's well-known argument that underlying mechanisms are irrelevant to the evolution of the male to female sex ratio in humans (Kitcher 1999) can

be adapted along the same lines. The vast majority of large animals have an approximately one-to-one sex ratio. This pleasingly even proportion is famously accounted for, in an explanation usually attributed to R. A. Fisher, as follows. The even sex ratio is a stable and unique evolutionary equilibrium. This is because, in a population with more females than males, individuals with a propensity to produce more males than females will have a higher expected number of grandchildren, and vice versa. Why is that? Your expected number of grandchildren is proportional to your expected number of children and your children's expected number of matings. Since matings require exactly one male and one female, a male's expected number of matings will increase, relative to a female's, as the proportion of males in a population decreases.

Although the physiological mechanisms and behavioral dispositions underlying mating, reproduction, and nurturance are different—often wildly different—in the various animals having a roughly even male to female ratio, it seems permissible, and even insightful, to say that the explanation of the ratio is the same in all animals that have it. The complete explanation of the ratio, if that is correct, black-boxes underlying mechanisms.[3]

(I should note that the version of the explanation given in the previous paragraph is not by anyone's standards complete. It is necessary, for example, to add that equal parental investment in the sexes is a precondition for the equilibrium—if it costs more to produce a male than a female, then the sex ratio will tend to be skewed toward females. But these additions will be black-boxing: the reasons for the equality of parental investment will not be spelled out or, at least, they will not be spelled out at a level of detail that distinguishes the different organisms to which the Fisherian explanatory scheme applies.)

3. The argument presented here departs from Kitcher's original argument in several ways. First, Kitcher's explanandum concerns humans only. Second, his explanandum is the fact that the ratio slightly favors males (because males are less likely to reach sexual maturity). Third, he compares the high-level explanation that black-boxes all facts of implementation with an ultra-specific explanation that recounts the conception and gestation of every human born over a certain period—a vastly more detailed explanatory story than any low-level evolutionary model seriously considered in this chapter.

## 4. Against Underlying Mechanisms

Let me develop the convergent evolution argument against the explanatory relevance of underlying mechanisms, arguing in favor of each of its two premises in turn.

The first and I think ultimately more contestable premise is that the explanation of the talpid phenotype is the same in golden moles and marsupial moles.

The talpid phenotype, I remind you, is defined so as to include only those features shared by golden moles and marsupial moles. There are many differences between the two taxa, including many differences in the way that the talpid phenotype is realized. Both kinds of organisms have very dense fur, but the patterns of fur growth are (let's say) not identical. To have the talpid phenotype is to have very dense fur, then, but it is not to have any particular pattern of fur growth. Consequently, an explanation of the phenotype should account for fur density, but it need not say anything about fur growth pattern. Indeed, it should not say anything about growth pattern, insofar as the pattern differs in the two kinds of mole. When explaining the golden moles' talpid phenotype, then, you are explaining the instantiation of *exactly the same property* as when you are explaining the marsupial moles' phenotype. The question is: do the two explanations nevertheless in some way differ?

Here is a simple argument for their not differing in any way. Both the golden moles and the marsupial moles have the talpid phenotype for exactly the same reasons, founded in the phenotype's adaptedness to the fossorial lifestyle—a burrowing, subterranean mode of existence. This is why it is a genuine case of convergent evolution. But the reasons that a taxon has a phenotype are just the explanation of that phenotype. So if the reasons for the phenotype are identical in golden and marsupial moles, the phenotype's explanation is in both groups identical.

Compare: the reason that a male cardinal's feathers are red is different from the reason that raven blood is red. So the explanations of redness are

different in each case; the double redness is coincidence rather than convergent evolution. There is no element of the talpid phenotype, by contrast, that is the same in golden and marsupial moles only by coincidence. Every aspect of that phenotype has its roots in the fossorial lifestyle.

On to the second premise of the argument, that the mechanisms underlying the evolutionary process in golden and marsupial moles are different in at least some relevant respects.

That the underlying mechanisms are different in some respects can hardly be denied. Every schoolchild knows that the mechanics of reproduction in marsupials are different from those in placental mammals.[4] Anyone with some biological sophistication can list many other differences in the causal underpinnings of survival and reproduction in these and for that matter in almost any two distantly related groups of organisms: different mating behaviors, different dentition, somewhat different diets, different numbers of chromosomes, and so on. (The latter two vary even among the different genera of golden moles.)

This in itself will not convince a proponent of the explanatory value of underlying mechanisms, however. No one, except perhaps a few extremists, believes that *every* aspect of underlying mechanisms is relevant to explanation. Paint a big rock black and hurl it at a window. The window breaks, but the black pigment, though it contributes to the weight of the rock and is perhaps the only thing to make direct contact with the window, does not play a part in explaining the breaking—whereas the rock's large mass, of course, does. An appealing way to separate explanatory from non-explanatory properties of low-level mechanisms is a difference-making account of relevance, according to which the rock's mass is relevant to the breaking and its paint job is not because the mass makes a difference to whether or not the window breaks and the paint makes no difference. But what follows does not turn essentially

4. Although as mentioned in the next section, not all marsupials have the eponymous pouch.

on any particular view of relevance.[5]

Here is the dialectical situation. Divide the philosophers with something to say about explanatory relevance, autonomy, and underlying mechanism into three classes.

First, there are those who hold that in the high-level sciences, underlying mechanisms are typically not explanatorily relevant. Complete high-level explanations, on this view, normally contain black boxes that stand in for all of the physical, or chemical, or (depending on the science) biological or psychological details. For these thinkers, the explanatory autonomy of, say, cognitive psychology from neuroscience is accounted for by the explanatory irrelevance, in psychology, of neural implementation.

Then there are those who hold (like me) that some details of underlying mechanisms—the difference-making details—are relevant in the high-level sciences, but are normally omitted in order to reap the efficiencies made possible by the division of explanatory labor. Confronted by the phenomenon of convergent evolution, and the case of the golden and marsupial moles in particular, they might react in one of two ways: they might acknowledge that the underlying mechanisms which differ between the two kinds of mole are on their view explanatorily relevant (this constitutes the second class of philosophers), or they might not (the third class). The former route means accepting premise (2) of the convergent evolution argument; a successful defense of the pragmatic approach to autonomy and black-boxing, at least in the case of evolutionary theory, then depends on finding some reason to reject premise (1). This strategy will be considered in the next section.

The third class of philosophers hold, then, that in general underlying mechanisms are explanatorily relevant and are ignored by the higher-level sciences only for practical reasons, but that convergent evolution, or perhaps evolution in general, is an exception. In particular, philosophers in this class

5. For a survey of ways to make sense of difference-making, including but not limited to counterfactual approaches, see Strevens (2008).

hold that in the case of the moles, the low-level mechanisms for survival and reproduction that distinguish the golden and marsupial moles are irrelevant to their talpid phenotype. When you have moles, then, neglect of lower-level mechanisms is explained by their objective explanatory relevance rather than by practical concerns. In general, however, this is not the rule; in general, underlying mechanisms are relevant and are ignored only to make science more efficient.

What are the prospects for the third kind of view? I am not aware of any account of explanatory relevance, whether based on difference-making or not, that will rule out the relevance of low-level mechanism *tout court* in evolutionary processes, or even only in subterranean evolutionary processes, yet that will attribute explanatory weight to low-level mechanisms in, say, psychology or economics. Consequently, I suspect that the third view is a case of adhockery in the service of wishful thinking. Maybe I am wrong; in the remainder of this chapter, however, I put the third view to one side, assuming that if the practical account of explanatory autonomy is to be defended against the argument from convergent evolution, it is premise (1) that must go.

## 5. On Behalf of Underlying Mechanisms

Newborn marsupial moles migrate to the mother's pouch, where they suckle in safety until maturity. Golden moles have no pouch; their young are more developed at birth and take shelter in a nest on their way to adulthood. The mechanisms underlying reproduction and the nurturing of young are therefore, in this and many other respects, different in golden and marsupial moles.

Are mechanisms like these explanatorily relevant in evolutionary theory? On one side of the question are philosophers like Kitcher who hold that implementational details of this sort are entirely irrelevant to models of certain high-level evolutionary processes, such as the evolution of the one-to-one sex ratio in humans and other animals and (I have suggested) the evolution of the talpid phenotype in the two groups of moles. On the other side are those like

me who hold that some details of implementation are explanatorily relevant in accounting for even the most abstract and high-level explananda, and that if they are routinely ignored, it is for practical reasons alone.

Cases of convergent evolution add weight to the anti-detail view: the explanation of the talpid phenotype is identical in golden and marsupial moles; thus, it cannot include aspects of their lives that differ, and so it cannot contain details about reproduction and nurture. Likewise, I have extrapolated, Kitcher would say that the explanation of the roughly one-to-one male to female ratio is identical in humans, moles, and many other creatures; the explanations in each case cannot, then, include the sexy details of reproduction, insofar as they differ from species to species.

This final section of the chapter will explore a two-part strategy for resisting such a conclusion, first giving a positive reason to think that the explanation of the talpid phenotype is slightly different in marsupial moles than in golden moles, and then giving an explanation for why we mistakenly think that they are the same. Let me emphasize that my aim in what follows is not to make a positive case for the explanatory relevance of underlying mechanisms; indeed, the suggested difference between the two explanations of talpid phenotype is not one of implementation. My tactics are purely defensive, then: I am trying to undermine the argument from convergent evolution, rather than to provide an independent, standalone argument for the relevance of implementation.

* * *

The pouch of marsupial moles faces backwards so that it does not scoop up sand and soil. Let me suppose for the sake of the argument that this orientation is essential; without the rear-facing pouch—that is, with a front-facing pouch or no pouch at all—marsupial moles could not sustain their fossorial lives. The evolution of the moles' talpid phenotype, then, required a rear-facing pouch either evolve (if the non-talpid ancestors lacked one) or that it be retained.

It seems that marsupial pouches have evolved, disappeared, and changed

configuration quite frequently during the time they have been around. Among the opossums of the Americas, for example, the pouch is usually absent or forward-facing, but in the yapok (*Chironectes minimus*), which unlike its mostly arboreal relatives forages underwater, it is backward-facing, presumably for much the same reasons as the marsupial mole pouch. Plausibly, then—though we are here in speculative territory—the marsupial moles' backward-facing pouch evolved at the same time as their talpid phenotype. Let me assume that it is so: the non-talpid ancestors either lacked or had forward-facing pouches, and in order to attain their fossorial lifestyle, the incipient marsupial moles had to evolve a backward-facing pouch along with the features they share with the golden moles.

In that case, it seems to me, the story of the development of the backward-facing pouch is an essential part of the story of the evolution of the marsupial moles' talpid phenotype, although the pouch and its orientation are not themselves a part of the phenotype. Why?

A proponent of the pouch's relevance might argue as follows. Relevance is a matter (so the speaker assumes) of counterfactual difference-making: *x*'s having *F* is relevant to its having *P* just in case, if *x* had not had *F*, then it would not have had *P*. If marsupial moles had not had backward-facing pouches, they would not have been able to adopt the fossorial lifestyle and so they would not have evolved the talpid phenotype. The pouch is therefore explanatorily relevant to the phenotype.

Such an argument is not decisive, however, because a defender of the argument from convergent evolution can make the following black-boxing reply. It is true that the marsupial moles would not have evolved the talpid phenotype if they had not had, or evolved, a reproductive system compatible with the fossorial lifestyle. But that is the right level to understand the relevance of the reproductive system: what mattered was *that* it was compatible with the lifestyle; the further details describing *how* it operated underground do not matter. The complete explanation of the talpid phenotype in marsupial

moles should black-box the details, then; it will specify only the fact of compatibility. The explanation of the phenotype in golden moles will of course specify precisely the same fact. The compatibility of the reproductive system is explanatorily essential, then, but it can be captured by a black box that sits equally easily in a specification of either marsupial mole or golden mole physiology.

Here is a better argument for the pouch's relevance: it is invidious to black-box the reasons for the reproductive system's adaptedness to the fossorial lifestyle (e.g., a backward facing pouch does not snag on dirt) while spelling out the reasons for the talpid phenotype's adaptedness to the lifestyle (e.g., dense fur enables the creature to slide easily past dirt). Since the explanation of the phenotype must describe the latter facts, it should describe the former facts as well.

My reason for thinking this is a certain explanatory holism about evolutionary history: the complete explanation of any of the marsupial mole's adaptations to life underground, I suggest, is an evolutionary story that relates all the important developments that made that life possible. These developments together make up a single evolutionary process; as they co-evolve, so they co-explain each other's evolution, because each next step in that evolution depends on the degree of fossorial compatibility so far attained. Or in other words, the springboard for the next step forward in the evolution of (say) the fur is in equal part the configurations of fur, claw, and pouch that enable the mole to dig a little deeper or a little faster.

This argument might, I think, be resisted by endorsing a principle according to which, when explaining the natural selection of a trait, you ought to black-box every aspect of the evolutionary process other than the mechanisms that constitute the trait. That is an extreme response: it would mean, for example, that there is no unified explanation of the talpid phenotype, but rather only a heavily black-boxed explanatory model for each component of the phenotype. I doubt that evolutionary biologists would recognize in this

explanatory atomization an accurate representation of their own scientific practice. But let me not pursue this line of thought here.

If I am right, the complete explanation of the talpid phenotype in marsupial moles contains an element—the causal history of certain properties of the pouch—that does not appear in the complete explanation of the same phenotype in golden moles. The two explanations are not identical, and the argument from convergent evolution therefore fails.

So what? After all, the difference between the two explanations is not exactly a matter of underlying mechanisms: the evolutionary history of the pouch does not lie at a lower level than the history of the talpid phenotype, and so its inclusion in the complete explanation of the phenotype does not constitute an explanatory descent to a lower level.

True, but it ought to lead you to draw a wider moral nevertheless. Positively, it shows that the evolution of the talpid phenotype is to be explained in part by specific facts about specific creatures, and so that the apparently desirable unified explanation of the phenotype, in the pursuit of which mechanism and many other particularities are jettisoned, is out of reach. Negatively, it suggests that the explanatory completeness of the unified, black-boxing explanation is in any case something of an illusion: we ought never to have thought that the story about the talpid phenotype was substantially the same in two kinds of animals that are so different deep down.

Let me now try to explain the source of that illusion.

* * *

There is something intuitively right about the claim that the talpid phenotype is identically explained in both the golden and the marsupial moles. I want to diagnose the source of that apparent rightness, and to show that it rests on a mistake.

The mistake is to conflate the complete explanation of why two things $x$ and $y$ share a certain property $P$, on the one hand, with the complete explanations of why $x$ has $P$ and why $y$ has $P$, on the other. As a result of

this conflation we infer, from the fact that $x$'s and $y$'s sharing $P$ has a unified complete explanation, that the complete explanation of $x$'s having $P$ is identical to the complete explanation of $y$'s having $P$.

I will be arguing, then, that the following "distributive principle" for explanation is false:

> Distributive principle: If the complete explanation why $x$ and $y$ both have $P$ is $M$, then the complete explanation of why $x$ has $P$ is $M$ (and likewise for $y$).

The sharing of a property by several entities is not only a different explanandum than the possession, by a single entity, of that same property; it is a different *kind of* explanandum.

Suppose that you are asked to explain why the US mens' basketball teams won the gold medal in the first seven Olympics in which basketball was played (1936–1968). You are to explain, then, why seven separate entities—the seven US teams—shared a certain property, namely, winning the Olympics.

The natural answer is something like this: basketball was established longer in, and was more popular in, the US than in any other country; the US in any case had a larger population than any other competitive country except the USSR; the US had in its college basketball tournaments a highly effective system for training young players; and so on.

These are properties that (presumably) have a role to play in explaining each one of the seven Olympic victories. But they hardly exhaust the factors relevant to the winning of any of the gold medals (or else the Americans would not have lost to Argentina in the 1950 FIBA championships, when these advantages also applied). Individual victories are explained by the skills of particular team members, the ability of those particular team members to work together in particular ways, and so on. As the composition of the team changed (because it consisted of college players, it was almost completely different for each Olympics), so these particular explainers changed. Should they not be a part of the story?

What the example shows, I think, is that the correct answer to an explanatory question about shared properties picks out only the explainers that are common to all of the relevant entities, that is, the factors that played a role in every one (or perhaps a majority?) of the wins. "What explains why *x* and *y* both have *P*?" is equivalent to "What factors appear both in the explanation of *x*'s having *P* and in the explanation of *y*'s having *P*?".

Further evidence for this interpretation is given by cases where there are no significant shared factors. Suppose someone asks why, in every US presidential election from 1952 to 1976, the Republican candidate won just in case an American League team won the (baseball) World Series. The answer is: there is no explanation, it is just a coincidence.

But clearly this series of events, like any other, can be explained. Indeed, presidential elections and major league baseball attract explainers like almost no other fixtures. You would not have to google far to find explanations for why Eisenhower won the election in 1952 and 1956, nor for why the (American League) Yankees won the World Series in those same years. Although there would be little or nothing in common between the baseball side and the presidential side of the account, then, you could give a perfectly good explanation of why things unfolded in the way that they did.

Why, then, say that the pattern has no explanation? Even if it is pure coincidence, it is hardly incomprehensible. The answer, as I have suggested, is that the explanatory request in cases like these is for shared properties in the explanations, that is, for factors that played a significant role in the causal histories of both the elections and the games of baseball. Perhaps the most important word here is "pattern": when asking for the explanation of a resemblance, a similarity, or some other run of events we want something that accounts not only for the individual events, but for their forming a pattern. That something will have to have figured over and over in the causal production of the instances of the pattern; when there is no such factor we say that the exhibition of the pattern (though not the facts that entail its

instantiation) has no explanation—it is just a coincidence.[6]

If what I have said is correct, then in explaining the similarities between golden and marsupial moles—thus, in explaining their convergently evolving the talpid phenotype—you will pick out only elements that play a role in explaining the phenotype's evolution in both groups. Thus, you will pick out the functional utility, for a burrower, of dense fur and spade-like claws, but not the importance of a backward-facing pouch, which is relevant only to the marsupials.

Incautiously, you might then apply the distributive principle spelled out above, and conclude that the features you have cited in your explanation of the shared phenotype also constitute the complete explanation of, first, golden moles having the phenotype, and, second, marsupial moles having the phenotype. That would lead you straight to premise (1) of the convergence argument, that the complete explanation of the phenotype is identical in both groups.

I suggest that this line of thought accounts for the appeal of premise (1). It is a mistake, however, because the distributive principle is false: a complete explanation of two entities' sharing a property is typically an incomplete explanation of each entity's instantiation of that property, as it leaves out by design explainers present only in one strand of the story, and thus aspects of the explanation of the dual instantiation that are "mere coincidence".

## 6. Conclusion

The high-level sciences that black-box most enthusiastically, and whose kinds are therefore the most promiscuously multiply realizable—economics, belief-desire psychology (if a science at all), mathematical ecology—can seem to be

6. These comments amount to an augmentation of the treatment in Strevens (2008, §5.5), where, though I noted the importance of citing similarities in explanations of similarities, I did not consider the possibility that the exhibition of a pattern over a given time period is a sui generis explanandum distinct from the facts that entail its instantiation. The difference between the two is, indeed, rather difficult to pin down, and I will not try to do so here.

alarmingly non-empirical in their content. In their characteristic explanations, what carries you from explainer to explanandum seems to be mathematical or logical rather than causal or physical necessity—or to put it another way, the phenomena predicted by these branches of science are represented as the consequences of theoretical definitions rather than of causal tendencies.

Is there a science of radically multiply realizable kinds that is plainly empirical, that identifies in these kinds real, explanatory causal tendencies rather than the logical black-boxy shadows thereof? Convergent evolution has long seemed to me to provide a promising testing ground for the idea that science can be radically multiply realizable but thoroughly empirical, thoroughly causal. It is in this role that I introduced it here, asking whether the explanation of convergent evolutionary tendencies dispenses with underlying mechanisms and other specifics, so as to offer a high-level causal unification of what are, from a low-level perspective, physically and indeed physiologically very different sorts of things.

Although my own views on explanation suggest that such a unification is impossible (Strevens 2008), I have ignored those views here and treated the case study in way that is as unladen with theory as I can manage.

I have provisionally concluded that even in a paradigm of convergent evolution, the marsupial and golden moles, there is not a genuinely unified explanation that crosses lower-level kinds: the explanation of the talpid phenotype in marsupial moles makes reference to properties that are not shared with the golden moles (and presumably vice versa).

This is hardly the end of the debate. I think that the argument for the splitting of the explanation along the boundaries of the lower-level kinds is strong but not decisive. Even if it is correct it does not follow that underlying mechanisms are typically relevant, or relevant even in this particular case. Radically multiply realizable kinds may have real causal-explanatory oomph in evolutionary and diverse other processes.[7]

7. For a theory on which black boxes can be causally explanatory, see Franklin-Hall

But I am somewhat pessimistic that progress can be made without debating the virtues and vices of particular philosophical accounts of scientific explanation. If I am right about the case of convergent evolution, then an impression of explanatory as well as evolutionary convergence is created by the rules for explaining shared properties; we will need theories of explanation to uncover and disentangle such intuitions.

Signing off, I leave things much as they were when I began. The high-level sciences are clearly black-boxing in their explanatory practices; they are autonomized if not atomized. But whether this segregation reflects the canons of explanatory relevance or merely a canny division of labor—whether, to return to the topic of this volume, explanations in cognitive psychology are independent entities existing quite independently of the details of neural implementation or whether they are explanatory sketches or templates awaiting neuroscientific substance—remains to be seen.

(forthcoming).

**References**


Cartwright, N. (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge University Press, Cambridge.

Dupré, J. (1993). *The Disorder of Things*. Harvard University Press, Cambridge, MA.

Fodor, J. A. (1974). Special sciences. *Synthese* 28:97–115.

Franklin-Hall, L. R. (Forthcoming). The causal economy account of scientific explanation. In C. K. Waters and J. Woodward (eds.), *Causation and Explanation in Biology*, volume 20 of *Minnesota Studies in the Philosophy of Science*. University of Minnesota Press, Minneapolis.

Garfinkel, A. (1981). *Forms of Explanation*. Yale University Press, New Haven, CT.

Hempel, C. G. (1965). Aspects of scientific explanation. In *Aspects of Scientific Explanation*, chap. 12, pp. 331–496. Free Press, New York.

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese* 183:339–272.

Kitcher, P. (1999). The hegemony of molecular biology. *Biology and Philosophy* 14:195–210.

Piccinini, G. and C. Craver. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese* 183:283–311.

Strevens, M. (2008). *Depth: An Account of Scientific Explanation*. Harvard University Press, Cambridge, MA.

———. (2016). Special science autonomy and the division of labor. In M. Couch and J. Pfeifer (eds.), *The Philosophy of Philip Kitcher*. Oxford University Press, Oxford.