Varieties of Explanatory Autonomy

Lawrence Sklar

Department of Philosophy, University of Michigan

This is the text of a talk given at the Robert and Sarah Boote Conference in Reductionism and Anti-Reductionism in Physics, 22-23 April, 2006, Center for Philosophy of Science University of Pittsburgh.

1

Everybody agrees that there are a multitude of scientific theories that are conceptually and explanatorily autonomous with respect to the fundamental concepts and fundamental explanations of foundational physical theories. Conceptual autonomy means that there is no plausible way to define the concepts of the autonomous theories in terms of the concepts that we use in our foundational physics. This is so even if we allow a rather liberal notion of "definition" so that concepts defined as limit cases of the applicability of the concepts of foundational physics are still considered definable. Explanatory autonomy means that there is no way of deriving the explanatory general principles, the laws, of the autonomous theory from the laws of foundational physics. Once again this is agreed to be the case even if we use a liberal notion of "derivability" for the laws so that derivations that invoke limiting procedures are still counted as derivations.

When talking about this conceptual and explanatory autonomy a theory is often said to be “irreducible” to foundational physics. More problematically we often hear talk of the autonomous theory revealing to us that phenomena in the world are “emergent” relative to the phenomena characterized by fundamental physics. It is probably best not to ask such questions as whether or not one theory is “irreducible” to another or whether or not some phenomena are “emergent” relative to other phenomena. The vagueness and slipperiness of these terms is unhelpful. Rather, our task is to recognize the wide variety of reasons why conceptual and explanatory autonomy can exist, and to carefully distinguish each such ground for autonomy from all the others.

2

The most dramatic kind of autonomy we could expect would be the case where the autonomous theory refers to a realm of being simply unaccounted for in our foundational physics. For a very long time this was the kind of autonomy that those who denied the universal reach of foundational physics really had in mind. More often than not they claimed to descry it in the realms of life and of mind.

Divine intervention, miracles, transcendent spiritual beings and the like – were they to exist – would presumably be events and things whose very ontology leaves them outside the realm of physical explanation. “Intelligent design” is just the latest sadly defensive version of such proposals.

But those more “naturalistically” inclined, and dubious of the supernatural in its entirety might still seek some sorts of “ontological” autonomy. Bergson and others posited the existence of “élan vital” or “life force” as the *sine qua non* of the possibility of biological life. Of course as our knowledge of the molecular biochemistry of biological phenomena increases at an exponential rate, the plausibility of such views, resting upon claims that no other explanatory account that left vital force out could possibly do, becomes less and less plausible. The one residual area where the purest kind of autonomy based upon a claim of a realm of ontology outside that comprehensible within physical theory is, of course, that of alleged directly accessible contents of sensory awareness. Whether we must countenance such things remains a great mystery. And whether their existence, if there be such, places any sort of limits on the universality of our physical explanations remains an open question. And let us put to the side any of the notorious attempts to show that some posited realm of the mental outside the realm of the physical must be taken account of in our most basic physical explanations when transcendent egos act on physical systems to collapse wave packets!

3

In recent philosophy explanatory autonomy has been treated in greatest detail in proposals dealing with the so-called special science. Put issues of sensory contents of mental awareness to the side. Indeed, grant to domain of

entities posited in fundamental physics full universality. Still a very plausible case can be made out that there is a plentitude of explanatory schemes, and schemes of concepts needed to frame these explanations, that is not derivable from, or, indeed, even closely related to the conceptual and explanatory repertoire of physics.

Economics treats of prices, capital flows and monetary equilibrium. Its explanations are framed in terms of market equilibria, rational expectations and the time value of money. The theory of human action treats of beliefs and desires and explains in terms of maximizing expected utility. Dynamic personality theories treat of attachment bonds, frustration and aggression and explain behavior in terms of psychodynamic development theories. What do such concepts have to do with those such as quantum relativistic fields or with such explanations as unitary time translations or measurement collapses?

Indeed, it has often been argued that biology is replete with its particular set of classificatory concepts and explanatory principles, and there is a substantial literature devoted to convincing us that the same thing is true of chemistry.

In each case we are given arguments (of varying plausibility depending on the cases) that any hope of “defining” the concepts of the special science in terms of the concepts of fundamental physics is out of the question. And we are given arguments (again of varying plausibility) that the explanatory principles of the special science “stand on their own” with no possibility and no need of being derivable in any way from the dynamical laws of physics.

The central role within some such theories of such principles as rationality of behavior (maximizing expected utility and the like), and the justifications of such principles out of fundamental posits such as preference transitivity for lottery choices, certainly shows a conceptual and explanatory structure at the heart of some of these theories that seems entirely independent of the casual and dynamical structure of fundamental physics. This, of course, is the modern day version of the traditional (Kantian) claim of autonomous realms of the rational (as opposed to the causal), but in an innocent version that can be claimed to require no mysterious posits of noumenal realms of acuasal freedom.

Actually the story of how the special sciences with their "autonomous" conceptual schemes and their "autonomous" explanatory principles relate to the conceptual framework and explanatory principles of fundamental physics is, ultimately, a very complicated tale. Some of it has been told but much more needs to be done. Too often we are presented with dramatic claims motivated by a kind of partisan demand for respect for a discipline as free of any need to bow to physics, and too little in the way of careful and detailed exploration of how the complicated web of the variety of sciences is actually constructed and how hierarchical elements (with physics at the bottom – or top? –of the hierarchy) play a crucial role in that construction.

Some basic aspects of this complicated structure are pretty easy to see, though. The disciplines that seem most remote from fundamental physics in their conceptual and explanatory structure are often those that explore some kind of "functional" order in a complex system. The system may be a society with

individual persons and their relations as the system whose functioning is in question. Or it might be components within a single functioning individual with a complex inner mechanism of parts whose interactions generate the behavior of the individual.

What do we need in the way of a contribution from physics to get a system that constitutes an economy? Only enough to ground a very abstract structure. We need notions of individuals in a society to be identifiable and reidentifiable over time. We need characterizations of interactions among individuals in terms of "transfers" of items abstractly characterized as "signals." So we must presuppose some kind of physical medium in which the abstractly represented transferred things or signals must be instantiated. Issues of a sufficiently complex inner nature of individuals to constitute a "memory" often play a role. Again we must presuppose that something in the physical makeup of the individuals can play this role, having itself component parts subject to change, persistent enough and stable enough to constitute "records" and things of that sort. But what the actual physical mechanisms are that play the role of the concrete realizations of the abstract components is of only the most marginal interest to the special science.

Hence, of course, all the philosophical fuss over "multiple realizability." A logic program can be realized in electron flows in silicon or water flows in a sufficiently complicated hydraulic system of tubes and valves. Carbon based humans form economic units, but we could well imagine (and sci-fi writers have) an economic system of individuals wildly differently constructed.

There are other cases, though, where the connection of the autonomous abstractly construed functional structure to the underlying physics (or at least chemistry) in which it is instantiated is much more to the fore. Evolutionary biology with its genes, genotypes and phenotypes and with fitness and natural selection as its guiding explanatory principles, could, of course, be realized in many different physical regimes. Indeed, the existence of such things as "evolutionary programming" where it is abstract programs (themselves realizable in multiple physical kinds of computers) that do the evolving, shows us just how broadly the evolutionary scheme can be applied.

But, of course, our deepest interest is in the evolution of the actual biological entities that constitute the realm of life on earth as that is that is the central concern of evolutionary theory in the first place. Now the identification of the transmissible unit of heredity, the gene, with the specific chemical constituent of the cell nucleus – the DNA – is all important. And with that comes much in train that explains why evolutionary theory in the abstract worked as well as it did, such as the now possible explanations of variation in trerms of mutation and chromosomal exchange mechanism and the like. And with it comes also all the insights that require major changes in the orthodox evolutionary theory, such as maternal imprinting for example.

In the case of chemistry, one becomes suspicious of claims to the effect that the autonomous concepts and autonomous explanatory schemes of this science are really that autonomous at all. Here the intimate connection between the chemical concepts and clearly identifiable underlying physical concepts, and

the close relationship of the chemical explanations to underlying patterns of physical explanation leads us immediately to think of the chemical concepts and explanations as somehow convenient ways of dealing with what is essentially physics and not anything like the functional social or even the biological sciences. I won't go into this now, but similar considerations will come to the surface shortly when we consider the issues of autonomy of concepts and explanations within physics.

4

Our main concern here is the issue of conceptual and explanatory *autonomy* within physics itself. What can this possibly mean? Well the natural understanding goes something like this: Our usual mode of explanation within physics is to characterize systems in terms of a limited vocabulary that picks out their basic dynamical states. This would be position and momentum for classical dynamics, basic field variables for classical field theories, perhaps the wave function for quantum dynamics and the like. Then we account for the behavior of systems by invoking the fundamental dynamical equations that link together basic dynamical states at different times (or, perhaps, at distinct spacetime locations).

But many of our explanations in physics don't look like that. Conceptual devices are invoked that don't seem to have any direct connection to the framework of the basic dynamical states. And novel explanatory patterns

emerge. Some of these seem to use lawlike connections that don't directly stem from the fundamental dynamical laws. Other explanations don't seem to have the standard dynamical linking of state to state as their structural framework at all. How are we to understand such "autonomous" pieces of physics itself?

The most famous of these cases of peculiarly autonomous bits of physics is, of course, thermodynamics. It is of absolutely fundamental importance within physics. It is amazingly universal. No matter what the basic entities you are dealing with and what their appropriate dynamics, the thermodynamical considerations apply. And it is strangely autonomous. Indeed, it would have to be in order to be applicable so independently of the constitution and fundamental dynamics of the systems to which it is applied. In particular temperature and entropy seem initially to have no place in the realm of the usual dynamical state concepts of physics. And the Second Law is, of course, the most notorious of the apparently autonomous explanatory principles of thermodynamics.

But, of course, thermodynamics is not as totally autonomous of dynamics as it first appears. Duhem and Mach were wrong, Maxwell and Boltzmann were right! Thermodynamics is situated in kinetic theory and statistical mechanics. From this perspective some concepts of thermodynamics and some of its basic laws seem very un-autonomous indeed. Heat flow is transfer of internal energy and the First Law is just the conservation of energy, which is a basic feature of dynamics and its time translation symmetry.

But what about the more puzzling concepts – temperature – and, notoriously, entropy? And what about the Second Law?

Much of foundational statistical mechanics is an exploration of the degree of autonomy from dynamics that remains to thermodynamics once the statistical mechanical approach is adopted. Here many questions still remain unanswered.

In equilibrium statistical mechanics ergodic theory can be viewed as seeking for the maximal degree of non-autonomy the statistical mechanic posits can be show to possess. The game is peculiar from the usual point of view of causal-dynamical explanations. We assume equilibrium states exist, without asking why or how they come into being. We use their time invariant nature to demand time invariance of the statistical mechanical account of them. Now comes the basic insight: We get the statistical mechanical account of the equilibrium situation by positing a probability distribution over the dynamical micro-states of the system. Then we calculate mean values of phase (dynamical) features and identify these with macro thermodynamic characteristics of systems. In all of this already many conceptual issues arise ranging from “Why mean values?” (partly answered by thermodynamic limit considerations that identify these with most probable values) to important questions about the relation of the macro to the probabilistic-micro descriptions (called “analogies” by the cautious Gibbs).

Ergodic theory tries to ground the posit that most distinctively suggests autonomy for statistical mechanics, a posit of probability over initial conditions of systems. From the dynamical perspective it doesn’t look like any such constraint on how initial states of a collection of systems ought to be distributed follows at all. But in the equilibrium case ergodic theory does tell us something about how

the usual probabilistic over initial conditions can be integrated with the underlying dynamical laws: Assume that the probability distribution over initial conditions is invariant in time – since, after all, we are trying to represent with the time invariant equilibrium state. Then ergodic theory tells us that, given the dynamics, if the system is "metrically indecomposable," then the usual probabilistic posit is the only one invariant in time that gives probability zero to collections of initial conditions given probability zero by the standard probability distribution. Of course we then need to show the systems we are dealing with are metrically indcomposable (crudely that they have no hidden global constants of motion), and that requires a whole world itself of idealization. And when we are done we still have two peculiar worries: First of all, all of this is done assuming equilibrium to be the case and is far from the kinds of causal-dynamical accounts we usually think of as explanations in our fundamental dynamics. And the "set of measure zero" problem remains glaringly autonomous from anything that seems to come from dynamics.

When we move to the non-equilibrium theory, the seeming autonomy of the needed posit of some probability distribution over initial conditions appears even more dramatic. We can hope to extract information from the underlying dynamics of the micro components in our search for the prediction of an approach to equilibrium and for a derivation of some kinetic equation describing that approach to equilibrium that can be connected by thermodynamic analogies to a macroscopic equation of evolution toward equilibrium. Indeed, it is dynamics when combined with facts about the constitution of our idealized systems (such

as being hard spheres in a box) that provides the resources for the various attempts at extracting a kinetic equation in non-equilibrium mechanics. Mixing as a generalization of ergodicity is one such approach. Topological approaches using such notions a topological entropy is another. Still a third is the Lanford "rigorous derivation of the Boltzmann equation" for a system suitably idealized to represent a dilute gas.

But no resort to dynamics and composition will do by themselves to ground the non-equilibrium theory. This is easily seen by the fact that both dynamics and composition are time-reversal invariant and the thermodynamics of non-equilibrium is not. The usual addition is, of course, some probabilistic posit over the initial conditions of the micro components at the time a system is isolated for study.

Here a number of considerations support the contention that something has been added to physics that goes beyond anything contained, even implicitly, in the structure of the underlying dynamics of the micro components. Some autonomous probabilistic posit must be made. And any hope of some a prior derivation of this from non physical principles is, despite repeated confused assertions to the contrary, out of the picture. Worse yet, in the non-equilibrium case we don't even have the "hemi-semi-demi-quasi" derivation of the probabilistic posit that ergodic theory provides for us in the equilibrium case.

An additional problem that hasn't received enough attention is this: Suppose we take for granted that we are to posit an initial probability distribution over the micro-states of a non-equilibrium time. And suppose that we accept the

usual posit that this distribution should be uniform over a characterization of the micro-dynamical states in a position-momentum phase space. We still need to characterize the region of the phase-space relevant to the problem at hand. This is generally done by relying on our knowledge of which array of macroscopic parameters is sufficient to characterize the macroscopic dynamical approach to equilibrium of the system. Sometimes reliance is placed upon the know macroscopic characterization of the equilibrium state of the system. In other cases one relies on known abilities to characterize the system in its non-equilibrium states and their changes (relying on such things as temperature and pressure fields, for example). This is then combined with the idea from the statistical mechanical side that the thermodynamic analogies tell us that the macroscopic features are associated with features of the micro-states by means of the *reduced* distribution functions. Density fields, for example, are derivable from one-particle distribution functions and macroscopic transport features from the two-particle correlation functions latent in the full probability distribution over microstates.

The crucial thing for us is the fact that we don't seem to have any way of picking out the "right" ensemble to use to characterize a system from first principles about its constitution and its dynamics, but rely instead on what we know about its ability to have its state and evolution characterizable in terms of macroscopic features experimentally available to us in the laboratory.

And on top of all of this is the notorious need to posit for the Big Bang initial state of the universe an astonishingly low entropy initial distribution, in the

form of uniform space or otherwise, in order to get first a time asymmetric approach to equilibrium for the cosmos as a whole and then, by some means not yet very clear, a parallel increase of entropy in the time direction in which the entropy of the whole is increasing for the “branch systems” without sneaking in additional time asymmetric probabilistic posits.

The case of thermodynamics and statistical mechanics does provide us, though, with something quite useful for our general methodological considerations about how concepts and explanations can be “autonomous” even within physics. There are many open questions about how to conceive correctly of the relation of temperature and entropy to features of the world characterized in the concepts of underlying fundamental dynamics and its states of systems. And there are certainly many open questions about how to fit thermodynamic explanations of the macroscopic behavior of things, and even statistical mechanical explanations produced to account for the thermodynamic regularities into the basic explanatory pattern framed by the fundamental laws of dynamics.

But at least we have some good idea of where the “autonomy” of thermodynamics and statistical mechanics comes from. Nothing in dynamics seems to constrain in any way how the initial conditions of a collection of systems will be distributed over the range of possibilities of those conditions allowable by the imposed constraints on the system. But statistical mechanics does demand such additional constraints, framed in terms of an imposed probability distribution on the initial conditions. Although we can find some degree of such constraints in some cases that stem from the underlying dynamics alone (ergodicity in the

equilibrium case, for example), it still seems as though we must posit additional constraints that go beyond anything given to us by the dynamical laws if we are to understand such bald facts about the world as the time asymmetric approach to equilibrium and its characterizability in macroscopic, thermodynamic terms.

5

Recently attention has been focused on the fact that *within* physics we find conceptual schemes and explanatory patters that seem to show a kind of autonomy relative to the usual foundational dynamical laws. This autonomy even persists if we count within the foundational concepts and laws those of thermodynamics and statistical mechanics. Bringing these realms of physics to the forefront of our attention has been invaluable. There can be no question that physics itself is constituted in important part by a variety of conceptualizations and explanatory modes that do not “reduce” in any simple-minded way to the concepts and explanatory patterns of the usual foundational physics.

There are phenomena that need to be explained that we are aware of from our macroscopic experience: the existence of dispersion phenomena stable under gross modifications of the dispersing medium, crystallization of solids, recurrence of features of phase changes over a wide variety of substances and phase change inducing transformations, and so on.

The explanations we are given for these phenomena are sometimes couched in principles derived not from the foundational theory thought to govern

what is going on, but, rather, from some older “phenomenological” theory whose status is taken to be at best “approximative” in the light of newer science. One example the essential use of geometric optics to deal with the rainbow’s stability over change of drop shape.

In other cases explanations are derived by clever applications of principles whose legitimacy stems from what we know empirically and phenomenologically about the situation in question. For example, we know that crystals have, in fact, correlations among the atoms composing the crystal that extend to arbitrary great distances. From this is derived the famous use of scaling principles in group renormalization arguments that allows us to derive essential aspects of the phase change leading to crystallization. The kind of unlimited correlational order hasn’t been derived from the underlying physics of the atoms, although, of course, that too comes into play in the explanations. From such considerations also come the methods that demonstrate to us that the “universality” of phase change features across a wide variety of substances (crystallizations, ferromagnetism, etc.) follows from a few basic features of the situation such as the dimensionality of the system in question, the number of degrees of freedom of its components, and a few restrictions on the forces governing the interactions of the atomic components (limited range, etc.).

Other ingenious and elegant explanations rely on dimensional analysis. From a basic understanding of which physical quantities ought to be taken as causally relevant to some phenomenon, much can be extracted about how they

are relevant in a specifically functional way merely from consideration of their basic dimensions (length, time, etc.) and that of the phenomenon in question.

In all of these cases the explanatory patterns differ sufficiently from the familiar one of deriving the results from some solution to the fundamental dynamical equations, or as a limiting result of some such parametrized solution, that a claim of conceptual and explanatory autonomy from the foundational dynamics for portions of physics has clear plausibility.

6

But what sort of "autonomy" is being exposed here? Certainly it isn't the sort that requires an extension of our ontology beyond that predicated by the underlying fundamental theories of the composition of systems and the dynamics governing the basic components. But there doesn't seem to be room here, either, for the kind of autonomy that thermodynamics and statistical mechanics has over the underlying dynamics either. That autonomy as we have noted rests on the special consideration that the thermodynamic theories introduce a new element into our basic physics not present in the foundational dynamical theory, the probability distribution imposed on the otherwise "free" range of initial conditions allowed to the systems in question. In the cases we have been looking at the explanatory schemes can plausibly be argued to be autonomous to an underlying foundational theory that *includes* the statistical element of foundational physics along with the basic dynamical theory.

One can imagine autonomy within physics that has other sources. For example, suppose that the behavior of a system depends not only on its internal constitution and the dynamics of those constituents, but on some special background state of affairs that imposes a regular “boundary condition” on the systems in question, but where the dynamics of the bounding system and even of its interaction with the system in question is not treated in any explicit way that would allow these to be considered in an approach to the problem that adverts to the foundational composition and dynamics. One can imagine such a situation as giving rise to an autonomous “special” physics whose application was limited to the situations where this implicit background boundary condition held.

But the kinds of autonomous explanations we are considering don’t seem to fit that pattern either.

I think the gut reaction of many philosophers and scientists looking at these patterns of autonomous explanation will be that somehow or another the explanatory results in question must already reside implicitly in the foundational dynamics of the constituents of the system. I don’t mean this simply in the ontological sense that whatever the systems do must be “supervenient” on the determined behavior of their components that is fixed by the underlying dynamics, but in the sense that there must be some way of deriving the explanatory pattern itself, in the intensional sense of a pattern “as we understand it,” from the underlying composition of the system and the dynamics of the components. The same intuition holds, I think, for the alleged autonomous explanations of chemistry noted earlier.

Now, admittedly, we are not going to find these explanations at the foundational level by looking at individuals solutions of the dynamical equations, or even at the limits of such solutions as some parameter changes. Where will we find them? Here the best I can do is arm-waving. There are many ways of exploring the fundamental equations. They have, for example, whole *spaces* of solutions and we can look at these. In one way, of course, that is what we do in statistical mechanics where we propose probability distributions over the possible solutions. But the spaces have topological aspects as well. Are certain kinds of solutions generic, for example? And the solution spaces can be explored for such issues as their stability under parameter changes. As I said, all I can do here is wave my arms.

As I have said, I think it is a major contribution to not only philosophy of physics but to methodological philosophy of science to bring to our attention the rich realm of conceptualizations and explanations in physics that do not fit the orthodox pattern of stipulating initial or boundary conditions and deriving a solution from the fundamental dynamical laws. What would be of tremendous value at this point would be some wide ranging exploration that would tell us, in general, what kinds of such autonomous explanatory patterns have been used in physics, and, more deeply, provide some understanding of how the types of such explanations can be understood to form a systematic family of explanatory schemes – if, indeed, they do.

And if the “gut intuition” is correct that these explanations must be comprehensible in terms of explanations that are formed at the level of the

concepts and resources of the fundamental theories of composition and dynamics, it would be wonderful to have some broad understanding of what the relevant explanatory features are at this foundational level and how in general they do their job of “grounding” the more phenomenological and autonomous explanatory structures.

7

But there are puzzling aspects to the claim that each of these autonomous modes of explanation within physics ought to be grounded on some non-autonomous structure extractable from the foundational equations.

After all, we all agree that disciplines such as economics, structural linguistics, personality dynamics, perhaps even abstract evolutionary theory all are genuinely explanatorily autonomous. We don’t demand any need to find a structure at the level of our universal, foundational physical theory that reveals at the deeper level the very explanatory structure of the special science. (Although we are happy in some cases, say evolutionary biology, when we can basically identify elements of the special science with elements characterizable at the more physical level – crudely, identifying genes with DNA molecules as the prime example.) And in all of these cases we don’t think that the claim for autonomy need rest on any ontological novelty or even on some special addition to the basic physics such as the probability distribution over initial states that makes thermodynamics and statistical mechanics so genuinely autonomous.

But for the physics cases the intuition is strong that not only do the foundational laws and the facts of composition ground the higher level structures and their regular behavior in the sense that the latter are supervenient on the former, but that the "purely physical" nature of the phenomena in question – structural stabilities, universalities such as in phase change, and the like, demand that with sufficient ingenuity an explanatory structure in the sense of something we grasp as informative that parallels the "autonomous" explanatory structures must be available at the foundational level.

But I really don't know how to characterize whatever good reasons for belief might lie behind such intuitions. Nor do I even know what should *count* as an explanatory scheme that is characterizable as being "at the foundational level." And, of course, we are a long way from understanding just what these foundational level structures of explanation ought to be for all of the cases of autonomous explanations within physics that have been so profitably brought to the attention of philosophy of physics.