STRUCTURAL EXPLANATION

ABSTRACT. This paper sketches an account of scientific explanation based on a semantic or model theoretic conception of scientific theories derivative from the work of Suppes and his collaborators [8, 10, 14] and a Bayesian account of scientific reasoning suggested by Rosenkrantz [9], Howson and Urbach [7] and Earman [3] among others. The model theoretic conception of scientific theories has been described in detail by Sneed [12], and Balzer, Moulines and Sneed [1]. A somewhat similar view – the semantic view – has been developed by van Fraassen [15], Suppe [13] and others. The discussion in this paper employs the conceptual apparatus developed in [1]. However, most of what is said can be applied to a more general ‘semantic’ conception of theories. The application of this conception of scientific theories to scientific explanation has been developed from somewhat different perspectives by Forge [4, 5] and Sintonen [11]. This paper generalizes (and somewhat simplifies) the work of Forge, indicates how it applies to so-called ‘functional explanations’, and connects it with the Bayesian account of scientific reasoning to provide an account of how competing scientific explanations are (should be) evaluated.

1. THEORIES AS MODEL CLASSES

On the model theoretic view, the simplest kind of scientific theory (theory element in the vocabulary of [1]) is an ordered pair \( T = (K, I) \) consisting of a conceptual core \( K \) and a range of intended applications \( I \). Very roughly, \( K \) defines a set of possible situations or ways things could be which we call

\[ \text{Content}(K). \]

The theory is used to say something – make a claim – about its intended applications \( I \). The claim is simply that \( I \) is one of those situations characterized by \( K \). Formally, this claim is just that

\[ (1) \quad I \in \text{Content}(K). \]

Note that formally \( \text{Content}(K) \) is a set of sets (of something). So that it makes sense to say that a set \( I \) is a member of \( \text{Content}(K) \). Roughly, we should think of \( I \) as the totality of potential data the theory
is supposed to account for. Individual members of $I$ are pieces of that
data – individual applications of the theory. Content($K$) characterizes
entire configurations of data – not just single individual applications.

In cases of genuine scientific interest, it appears that $I$ will be an
intentionally described, ‘open-ended’ class – for example the set of all
particle collisions – the set of all state-level societies. In these cases,
though we may have good reasons to believe $I$ to be finite, we will never
be certain that we have examined all its members. Thus, however sure
we may be of the truth of (1), there exists the possibility of discovering
additional members of $I$ which would cause us to revise our epistemic
attitude toward (1). Indeed, I think one might persuasively argue that a
necessary condition on ‘real’ science is that $I$ be ‘open-ended’ in this
way.

It is perhaps worth noting that (1) may be recast into the form of a
‘general statement’:

(1’) For all $X$, if $X$ is an $I$ then $X$ is a Content($K$).

In this form the claim of a theory looks very much like the ‘general laws’
that play a major role in traditional accounts of scientific explanation.
Roughly, what makes claims like (1), (1’) ‘interesting’ is that they point
out some feature that all members of $I$ – both those we have examined
and those we have not – allegedly have. The major differences between
the account of explanation I will offer and the traditional account is that
my account makes explicit use of the set-theoretic properties of things
that are (in) $I$ and things that are (in) Content($K$). The ‘features’ that
members of $I$ allegedly have are ‘structural’ features most naturally
described by the language of set theory.

Again, in cases of genuine scientific interest, Content($K$) has the
formal property that, whenever $Y$ is in Content($K$), so is every non-
void subset of $Y$. Intuitively, this means that inductive evidence for
(1) may be provided by examining subsets of $I$ and discovering them
to be in Content($K$). Indeed, one, overly simple picture of ‘theory
confirmation’ is that we just examine more and more members of $I$. If
the set of those we have examined continues to be in Content($K$), we
become more and more confident that (1) is true.

It is convenient to view this simple picture of theory confirmation as
a ‘research program’ carried out over time. Thus, at any time $t$ during
the course of this program there will be some subset $F_t$ of $I$ that is
‘believed with good reason’ (by members of the scientific community
pursuing the research program) to be in Content($K$). We may call $F_t$ the set of ‘firm’ applications of $T$ at time $t$. Firm applications that are in fact in Content($K$) we may call ‘successful’ applications. On the customary account of ‘knowledge’, the users of $T$ know that successful applications are in Content($K$). (These ideas are elaborated in detail in [1], Ch. V.) One may provide an account of explanation based either on ‘firm’ or ‘successful’ applications. The difference is just in the epistemic strength of the notion of explanation. To make matters simple, I will opt for ‘successful’ and an epistemically strong notion of explanation.

So far, very little has been said about the explicitly model theoretic nature of this conception of scientific theories. This becomes apparent only when we characterize $K$ in more detail. Intuitively, $K$ consists of a vocabulary or conceptual apparatus characteristic of the theory ($M_p$) from which a non-theoretical ($M_{pp}$) may be distinguished, some laws ($M$) formulated in the full vocabulary and some constraints ($C$) which limit the ways in which theoretical concepts may be applied across different applications of the theory. Thus $K$ is an ordered 4-tuple:

$$K = \langle M_p, M_{pp}, M, C \rangle.$$  

(2) 

What makes this conception ‘model theoretic’ is that each of the elements of $K$ may be viewed as a class of models – set theoretic structures. These model classes may sometimes be characterized by sentences in first-order logic. But, for theories of genuine scientific interest, characterizing them in the language of set theory by defining a set theoretic predicate is almost always easier and sometimes the only practical means. The insight that classes of models characterized in this way could serve to illuminate parts of interesting empirical science is due to Suppes and his collaborators [8, 10, 14].

These model classes hang together in the following way. Non-theoretical structures – members of $M_{pp}$ – are obtainable from members of $M_p$ simply by lopping off their theoretical components. Formally, there is a ‘forgetful functor’ that maps $M_p$ onto $M_{pp}$. $M$ is just a subset of $M_p$. It represents scientific laws in that this subset is a characteristic way in which the ‘values’ of the components of single members of $M_p$ are related. $C$ is a subset of the power set of $M_p$ ($\text{Pot}(M_{pp})$). It represents characteristic ways in which the values of the components in different members of $M_p$ are related.

Content($K$) consists of all the sets $N$ of non-theoretical structures (subsets of $M_{pp}$) that can be filled out with theoretical components in a
way so that: (i) each individual member of $N$ is filled out to a member of $M$ (satisfies the laws); and (ii) the set of theoretical structures produced by filling out every member of $N$ is a member of $C$ (satisfies the constraints).

The set of intended applications $I$ is a set of non-theoretical structures described in the manner sketched above. Thus the claims of the theory (1) amounts to saying that $I$ (and every subset of $I$) can be filled out with theoretical components in some way that satisfies these conditions. That this rendition of the empirical claim of theories like classical particle mechanics, classical equilibrium thermodynamics and others is plausible has been argued in detail in [12, 1]. These arguments will not be repeated here.

Somewhat more complicated kinds of scientific theories can be represented as 'nets' of theories of the sort described here (theory elements) and an appropriate concept or 'content' can be defined for these nets. Sintonen [11] has exploited this more general concept of scientific theory to explicate some contextual, pragmatic aspects of the concept of scientific explanation. This macroscopic, net-level account of explanation effectively presupposes a microscopic, theory-element-level account which is not explicitly provided. This paper complements the work of Sintonen by providing the microscopic account it presupposes.

2. SCIENTIFIC EXPLANATION

These ideas may be applied to scientific explanation in the following way. Very roughly, the occurrence of $X$ is explained by showing it to be a specific part of something else that regularly occurs in a well-understood way. For example, the occurrence of long-neck, brown bottles in the recycling bin in front of my house is explained by showing it to be a part of a regularly occurring pattern of beverage consumption and refuse disposal characteristic of my household.

A bit more precisely, to explain some phenomenon $X$ is to show that $X$ is a specific aspect or part of some larger phenomenon sharing some interesting features with (generally occurring in the same ways as) a wider class of phenomena. To recall Hempel's classic example, we explain the bursting of my pipes last December 28 by noting that my pipes are part of a thermodynamic system and thermodynamic systems behave in certain general ways described in detail by a physical theory like classical equilibrium thermodynamics. We explain the entry of
the 8-ball into the corner pocket by noting the context in which this event occurred – a collision sequence involving the 8-ball, the cue-ball and a third ball. Further, this sequence of collisions has some general properties characterized by classical collision mechanics. We explain the presence of turquoise fragments on part of the surface of a grooved stone by noting features of the context in which the stone was found. It was in a stone enclosure on a river bank about a mile from a turquoise outcropping. We conclude the stone was part of a certain kind of ‘technological system’ – perhaps a ‘mineral extractive system’. Further, we are able to say something in general about ‘mineral extractive systems’. That is, we have a ‘theory’ about them.

Though these rough ‘explanation sketches’ do not make it explicit, it is worth noting the somewhat obvious fact that the intuitive plausibility of these explanations depends on identifying the phenomenon to be explained with a specific kind of part of the larger system. Though it is true, it is not enough to say the turquoise fragments were just any part of a mineral extractive system. We must specifically identify them as debris resulting from the use of a specific kind of tool used to break turquoise bearing rock.

This rough, intuitive gloss of ‘explanation’ does not sound too different (I hope) from more traditional accounts. Surely, it should not. My intention is to capture the same intuitive idea as the traditional accounts. The difference between my account and the traditional accounts lies in the way these initial intuitions are made precise. Note too that these examples are prima facie mute on the traditionally controversial issue of whether (a description of) the explained phenomenon is logically entailed by the explanation.

What all these examples have in common is that a phenomenon \( X \) is brought within the scope of some plausible theory. The epistemic value of explanation is that it allows us to view \( X \) as an aspect of ‘the familiar’ – things we think we understand. To begin to make this account more precise, we might say roughly this. To explain some phenomenon \( X \) is to show that \( X \) ‘is’ (may be viewed as) a part of some known, successful applications of a theory that can be taken seriously. On this account, providing an explanation for \( X \) can be analyzed into three parts:

(A) Showing \( X \) is a specific part of some \( A \);
(B) Showing \( A \) are successful applications of some theory \( T \);
(C) Showing \( T \) can be taken seriously.
Let us postpone saying what it means to say a theory can be taken seriously. I will return to this aspect of explanation in Sections 4 and 5 below. Rather, I will focus on (A) and (B) – explaining what it means to say that X is brought within the scope of theory T.

The phenomenon X is explained by bringing it within the scope of theory \( T = \langle K, I \rangle \). This simply means that we identify X as a specific part of some intended applications \( A \) of \( T \) that are in fact successful applications of \( T \). Thus we explain why the 8-ball went in the corner pocket by describing a collision of which its path to the corner pocket was a part and noting that this collision is a successful application of classical collision mechanics – i.e. it conserves momentum without requiring us to assign masses to the billiard balls that are incompatible with data about other collisions in which they have been observed. In this example X is shown to be a part of a single successful application. This is perhaps the typical case. However, the suggested formulation also allows for the possibility that X is shown to be a part of a set of successful applications. For example, X might be data about the role of a single particle – the cue ball – in several different collisions.

Note that ‘part’ here is to be understood to include ‘improper part’. Thus, showing that \( X \) is (all of) some successful intended application counts as explaining \( X \). In particular, showing that \( X \) is an intended application that can be added to the set of unsuccessful applications (at time \( t \)) without the resulting set falling outside of Content(\( K \)) counts as explaining \( X \). This appears to be the case considered by Sintonen [11] for theory nets. This kind of explanation is ‘progress’ of the research program associated with the theory. But there are more mundane kinds of explanation – perhaps those involving technological applications of the theory – that are not cases of ‘progress’ in any interesting way. The account offered here includes both cases.

Very roughly we have something like this:

(3) \( T = \langle K, I \rangle \) explains \( X \) iff \( X \) is a specific part of a set of some successful intended applications of \( T \).

Intuitively, a ‘part of’ an application is just something that can be filled out to a (full) application. For example, it might be a specification of the velocities of a subset of the particles involved in a collision; it might be the specification of the velocities of all the particles before, but not after, the collision. We need to talk about parts of sets of structures and this may be done in roughly the following way.
(4) \( X \) is part of \( A \) iff each member of \( X \) is part of exactly one member of \( A \) and each member of \( A \) contains exactly one member of \( X \).

Formally, a ‘part of’ an individual set-theoretic structure is just a substructure. A set-theoretic structure has lots of different substructures – different parts of the structure. An explanation of \( X \) must specify which specific substructure is identified with \( X \). For present purpose, I will use the notation ‘part-i’ to denote a specific kind of substructure of intended applications of \( T \). In doing this, I am ignoring certain technical questions about how substructures of intended applications are specified. Thus, a slightly more precise version of (3) is:

(3') \( T = \langle K, I \rangle \) explains \( X \) iff there are some \( i \) such that \( X \) is part-i of a set of some successful applications of \( T \).

We now need to be a little more explicit about the pragmatic aspects of explanation. Successful applications of a theory \( \langle K, I \rangle \) at time \( t \) are simply intended applications that are known by the users of \( T \) to at time \( t \) to be in \( \text{Content}(K) \). Thus:

(5) \( A \) is a set of successful applications of theory \( T = \langle K, I \rangle \) at time \( t \) iff:

(i) \( A \subseteq I \);
(ii) \( A \in \text{Content}(K) \);
(iii) at time \( t \) users of \( T \) have good reason to believe that \( A \in \text{Content}(K) \).

Thus, we have the following account or definition of ‘explanation’:

(6) \( \langle K, I, A \rangle \) is an explanation for \( X \) at time \( t \) iff:

(i) \( \langle K, I \rangle \) is a theory element;
(ii) \( A \subseteq I \);
(iii) \( A \in \text{Content}(K) \);
(iv) at time \( t \) users of \( T \) have good reason to believe that \( A \in \text{Content}(K) \);
(v) \( X \) is part-i of \( A \).

This account of explanation is not free from known problems. These turn on the concept of ‘part’ that is appropriate here. Roughly, some ‘parts’ of structures seem to be inappropriate candidates for explanation according to this model. For example, one might take a ‘part’ of a classical collision mechanical structure to be simply a subset of the set of particles appearing in the structure. It seems counterintuitive to say
that we 'explain' this set of particles by embedding it in a model for classical collision mechanics. The obvious solution to this problem is to define 'part' in such a way that subsets of non-basic sets in structures must appear in 'parts' of these structures.

Another problem appears in connection with theories (like classical collision mechanics) which have a concept of 'temporal order'. Suppose we 'explain' a given structure by identifying it as that part of a classical collision mechanical structure consisting of (some of) the particles plus their initial velocities. One might contend that doing this explains nothing about 'why the particles have these velocities'. In contrast, identifying parts of the same given structure with the final velocities of some other models of collision mechanics does seem to provide an intuitively satisfactory explanation.

Roughly, the same data may be explained by classical collision mechanics in several ways. Some of these ways appear to be intuitively satisfactory. Some do not. The source of this difference clearly lies in our intuitions about temporal order. In the intuitively unsatisfactory explanations we appear to be explaining the given structure by showing 'what happens next' rather than 'what happened before'. We might avoid this by placing additional restrictions on the temporal order of 'parts' appearing in explanations. Those who think that 'causes' are an essential feature of explanation would endorse this move. Thought this might be a feasible move for theories with a concept of temporal order, it is hard to see what (if anything) might correspond to this for theories (like equilibrium thermodynamics) without a concept of temporal order.

Perhaps the account of explanation offered here should be viewed as characterizing a rather general framework for explanation which can, in some cases, be augmented with further conditions to characterize 'causal explanations'. However, it should be noted that this augmentation will apparently usually involve adding something that is not naturally regarded as a part of the conceptual apparatus of the theory providing the explanation. That is, 'causality' – even when we can make sense of it – is usually not going to appear explicitly in the conceptual apparatus that is captured by the set-theoretic axiomatization.

Having said what I can about known difficulties, I will explore implications of this account of explanation in two directions. First, I will consider the traditional question about the symmetry of explanation and prediction in relation to the question of functional explanations.
Second, I will consider how comparative evaluations of explanations are made.

3. EXPLANATION AS ARGUMENT

Traditional accounts of explanation have tended to view an explanation as a kind of argument. Rough, the conclusion of the argument is that which is explained while the premises are provided by the theory that provides the explanation, together with some 'collateral information' or 'initial conditions'. How can we compare the present account of explanation with this 'deductive model' of explanation?

The most obvious approach to comparison is trying to convert the various pieces of a model theoretic explanation into linguistic entities that at least could be pieces of a deductive argument. For the theoretical part of the explanation, the obvious candidate is simply the claim of the theory \( T = \langle K, I \rangle \) that

\[
I \in \text{Content}(K).
\]

It is also reasonably clear that \( A \) – the set of intended applications used in the model theoretic explanation – plays a role corresponding roughly to that of the collateral information in the deductive model.

A bit more precisely, what remains of the set-theoretic structures in \( A \) when the explanandum \( X \) is deleted – this ‘\( A\{\ldots\} X \)’ is the collateral information. More intuitively, \( A \) is the total context within which we locate \( X \) for purposes of bringing \( T \) to bear on it. \( A\{\ldots\} X \) is that part of the context which was not initially given, but rather added (perhaps discovered) in the process of providing the explanation for \( X \).

Thus, to provide additional premises for a deductive argument we need to assert that the collateral information \( A\{\ldots\} X \) is, in fact, part of an intended application of \( T \). This we may do with two propositions:

\[
(7) \quad A \subseteq I
\]

and

\[
(8) \quad A\{\ldots\} X \text{ is part of } A.
\]

With (8) we simply claim that the additional structure \( A\{\ldots\} X \) which we have added to \( X \) is a part of the total context. This is, of course,
trivially true. Non-trivial is the claim (15) that the total context $A$ is an intended application of $T$.

What is now the appropriate conclusion to this argument? I suggest that we would like to be able to conclude from (1), (7), and (8) that:

(9) $X$ is part of $A$.

That is, we would like to conclude from the truth of the theory's claim; the fact that the theory applies to the total context $A$ and the fact that the added context $A\{\neg\}X$ is a part of $A$ that $X$ is a part of $A$.

Thus, I suggest that the most natural way to convert a model theoretic explanation into a deductive argument yields the following argument:

(1) $I \in \text{Content}(K)$.  

[T is true.]

(7) $A \subseteq I$.  

[T applies to $A$.]

(8) $A\{\neg\}X$ is part of $A$  

[collateral information]

(9) $X$ is part of $A$  

[explanandum]

The problem with this argument form is that it is not generally valid. Whether it is valid depends on the nature of the theory $T$ used in the explanation and upon the nature of the application $A$ of the theory. The reason is roughly this. For applications $A$ in $\text{Content}(K)$, parts of $A$ having the structure characteristic of $A\{\neg\}X$ will not always be associated with unique parts of $A$ having the structure characteristic of the part $X$. Thus simply knowing how the $A\{\neg\}X$-like part is instantiated does not generally tell us how the $X$-like part is instantiated.

To make this clearer, consider the special case in which $K$ contains no theoretical concepts and trivial constraints. That is,

\[ M_{pp} = M_p; \quad C = \text{Pot}(M_p). \]

In this case,

\[ \text{Content}(K) = \text{Pot}(M) \]
and nothing is lost by simply forgetting about multiple applications and taking

$$\text{Content}(K) = M.$$ 

Suppose that the explanandum $x$ is a substructure of some $m_p$ in $M_p$. That is, $x$ is obtained from $m_p$ by deleting some parts of $m_p$. Intuitively, we may think of systematically going through all of $M_p$ removing from each member parts isomorphic to those we removed from $m_p$ to obtain $x$. Call the set of structures we obtain in this way $S_x[M_p]$. Providing a precise definition of $S_x[M_p]$ is somewhat tedious so I omit it here. Intuitively, $S_x[M_p]$ is the set of all $x$-type data structures. The structure $x$ is just one ‘value’ or ‘instantiation’ of this data type.

To explain $x$ is simply to add the remaining structure required to make it an $a$ in $M_p$ which is also a model for $T$, i.e. $a$ in $M$. This remaining structure is $a\{\}x$. As before, we may consider $S_a\{\}x[M_p]$ – intuitively the set of all $a\{\}x$-type data structures.

Now, starting with

$$a\{\}x \in S_a\{\}x[M_p]$$

we may generally add the remaining $x$-type data to produce a model for $T$ in a variety of ways. That is the emendations of $a\{\}x$ in $M$ form a ‘blob’ rather than a point (see Figure 1). In turn, this blob of emendations to models for $T$ projects down into $S_x[M_p]$ as a blob, rather than a point. In this blob are all the possible ways of filling out $a\{\}x$ with $x$-type data that yields a model for $T$.

Thus, knowing only that $a\{\}x$ is a part of a model for $T$, the most we can infer about the value of the $x$-type part of this model is that it is somewhere in the blob projected into $S_x[M_p]$. In some special cases, this blob will be a singleton. But generally it will not be. Whether or not there is a unique way of making $a\{\}x$ into a model for $T$ depends on the specific nature of $T$ and $a\{\}x$.

Thus structural explanations are, in general, what Hempel called ‘partial explanations’ ([6], pp. 16–17). They allow us to deduce some features of the explanandum. But, they do not, in general, allow us to deduce everything we know about it – not even everything we know about it that is expressible in the vocabulary of the explaining theory. What the structural account of explanation allows us to see (which is, at least, not readily apparent in linguistic accounts) is that partial
Fig. 1.

Explanations occur in the 'hard' sciences as well as the 'soft'. Physical theory offers partial explanations that are, in their logical form, no different than the functional explanations appearing in biology and cultural anthropology. The only difference is this. In some cases, at least, physical theories can provide 'complete' explanations. It might be the case that the explaining theories used in sciences like biology and anthropology are simply so weak that the functional explanations they provide can never be more than partial explanations. Indeed, one account of functional explanation suggests that this is true [16]. If theories supporting functional explanations are 'optimization models' with multiple local optima, then it is intuitively clear that the explanations will always be partial.
4. BAYESIAN RATIONALITY

The obvious way to apply a Bayesian account of scientific reasoning to a model theoretic conception of scientific theory is to consider conditional probabilities of the following form:

\[
P(I \in \text{Content}(K) \mid E)
\]

where \(E\) is the ‘total evidence’ available. \(E\) might simply be a conjunction of things like

\[I_i \in \text{Content}(K)\]

\[\text{not } (I_j \in \text{Content}(K))\]

and

\[P(I_i \subset I) \neq 0.\]

But there is no reason at this point to rule out other kinds of evidence.

5. EVALUATING EXPLANATIONS

It is common practice among empirical scientists (and others) to debate the relative merits of different explanations offered for the same phenomenon. Do we best explain an apparent rise in temperature in a palladium cell by the presence of ‘cold fusion’ or by ‘faulty instrumentation’. Do we best explain the construction of large masonry structures, roads, etc. centered around Chaco Canyon, NM in the 10th–11th Century AD by some variant of ‘indigenous development’ or some variant of ‘foreign influence’. At least a half dozen alternative explanations for this phenomenon are evaluated comparatively in a recent discussion ([11], p. 391–402). Supposing that the alternative explanations in these examples could be mashed into the format suggested above, would this illuminate the discussion of the alternatives?

Given that we have two distinct putative explanations for \(X\), how should we compare them? Intuitively, two considerations appear to be relevant: (1) how good is the theory involved in the explanation?; (2) how probable is it that the theory involved really does explain \(X\)? On
a Bayesian account of scientific reasoning this suggests that something like:

\[
P(I \in \text{Content}(K) \& \langle K, I, A \rangle \text{ is an explanation for } X \mid E)\]

could be taken as a ‘figure of merit’ for explanations. That is, alternative explanations of \(X, \langle K, I, A \rangle\) and \(\langle K', I', A' \rangle\), should be evaluated according to the value they give for (11). In fact, it appears that in many cases of interest this expression can be considerably simplified.

First, it appears that in many cases of interest, the conjuncts in (11) will be probabilistic independent so that (11) reduces to:

\[
P(I \in \text{Content}(K) \mid E) \times P \quad (\langle K, I, A \rangle \text{ is an explanation for } X \mid E).
\]

Clearly this independence assumption will not generally be true. In some cases that \(\langle K, I, A \rangle\) explains \(X\) may provide evidence that \(I \in \text{Content}(K)\), that is

\[
P(I \in \text{Content}(X) \mid \langle K, I, A \rangle \text{ explains } X \& E) > P(I \in \text{Content}(X) \mid E).
\]

Intuitively, this is because

\[
A \subseteq I \quad \text{and} \quad A \in \text{Content}(K)
\]

provide evidence that

\[
I \in \text{Content}(K).
\]

However, in the case of theories with lots of evidence of this sort, the contribution of the evidence from specific putative explanation may be negligible. So the product form may be acceptable as an approximation. This would most likely be the case when the theories were well established and supported by substantial amounts of evidence apart from the explanatory context in question.

The product form (12) has the nice intuitive feature of clearly separating judgements about the merits of the theory invoked in explanation from judgements about whether the theory explains the phenomenon in question. Focusing on the latter judgement, some additional simplification is possible. Considering (6) again, one might most naturally view (6-i), (6-ii), and (6-iii) as purely formal conditions so the probability
of each (and the conjunction) will always be either 1 or 0. We could complicate things here by considering non-trivial probabilities for the formal conditions, as for example when we are uncertain about our mathematical or calculational results. But this complication would add little to the present discussion. Thus in the case the formal conditions are satisfied \( (P(6-\text{i}) \& 6-\text{ii}) \& 6-\text{iii}) = 1 \), the probability that we have an explanation at all is just the probability that the empirical condition (6-iv) is met. Thus, the probability that \( \langle K, I, A \rangle \) explains \( X \) is:

\[
(13) \quad P(X \text{ part of } A \mid E).
\]

Thus far we have avoided explicit discussion of the formal properties of the explanandum \( X \). In some cases \( X \) may be (described as) a set-theoretic structure in a fragment of the vocabulary of the theory \( T = \langle K, I \rangle \) used in the explanation. In this case, whether \( X \) is a part of \( A \) may be a purely formal question. In the more general case \( X \) will be described in some vocabulary disjoint from that of \( T \) and it will remain a ‘contingent’ question whether this description of \( X \) has been appropriately ‘translated’ into the vocabulary of \( T \).

Under these simplifying assumptions, the goodness of the theory is measured by the probability (given the available evidence \( E \)) of the empirical claims of the theory, i.e.

\[
P(I \in \text{Content}(K) \mid E).
\]

The probability that the theory explains \( X \) is:

\[
P(X \text{ part of } A),
\]

and the ‘figure of merit’ for explanations is simply the product of these probabilities. Thus:

\[
(14) \quad \text{If } \langle K, I, A \rangle \text{ and } \langle K', I', A' \rangle \text{ are putative explanations for } X \text{ then:}
\]

\[
\langle K, I, A \rangle \text{ is better than } \langle K', I', A' \rangle
\]

iff

\[
P(I \in \text{Content}(K) \mid E) \times P(X \text{ part of } A \mid E)
\]

\[
> P(I' \in \text{Content}(K') \mid E) \times P(X \text{ part of } A' \mid E)
\]
Thus far we have assumed that the theories involved in alternative explanations are totally different. That is, both their vocabulary and laws are different. This is not always so. In some cases of interest, the alternative theories will share all or part of their vocabulary. The most interesting of these cases are those in which the theories in question are "competing" theories about the same data – that is, when $M_{pp} = M'_{pp}$ and $I = I'$. In this case, the conditional probability of the theories' claims (10) do not always provide a good indication of the relative merit of the theories. The reason is roughly this: relatively weak theories may have claims with high probabilities. But, intuitively, stronger theories whose claims have lower probabilities might be preferable. How should theories be compared in these cases? And once theories are compared, how should explanations based on them be compared?

From a model theoretic point of view, the content of theories provides an indication of their relative strength. Clearly,

\[(15) \ T = \langle K, I \rangle \text{ is stronger than } T' = \langle K', I \rangle \]

if

\[\text{Content}(K) \subseteq \text{Content}(K')\]

whether the 'if' in (15) should be 'if and only if' is not entirely clear. More explicitly, it is not evident how we should compare the strength of theories with identical classes of non-theoretical structures whose contents intersect partly or fail to intersect at all. For present purposes, it appears that little is lost by simply taking (15) to be necessary, as well as a sufficient, condition. We may, somewhat arbitrarily, say $T$ and $T'$ are equally strong (equipotent) in all other cases. Thus:

\[(16-i) \ T = \langle K, I \rangle \text{ is stronger than } T' = \langle K', I \rangle \]

\[\text{iff} \quad \text{Content}(K) \subseteq \text{Content}(K')\]

\[(16-ii) \ T \text{ is equipotent with } T' \quad \text{iff} \]

\[T \text{ is stronger than } T' \text{ and } T' \text{ is not stronger than } T.\]

Having a way of comparing the strength of theories allows us to address the question of how probability trades off against strength in evaluating
theories. I suggest that the trade off is made lexicographically with probability dominating strength. That is, strictly more probable theories are always preferred to strictly less probable. Only in the case of equally probable theories does strength play role. Then, stronger theories are preferred to weaker. Thus:

(17) $T = \langle K, I \rangle$ is better than $T' = \langle K', I \rangle$
	niff

$$P(I \in \text{Content}(K) \mid E) > P(I \in \text{Content}(K') \mid E)$$

or

$$P(I \in \text{Content}(K) \mid E) = P(I \in \text{Content}(K') \mid E)$$

and

$T$ is stronger than $T'$

This seems plausible for the following reason. Numerous weak theories with high probability are easy to construct. But most such theories receive short-lived attention because it is also relatively easy to strengthen some of the theories without substantially diminishing their probability. The theories we consider seriously are those which retain their probability under strengthening. Having a ‘better than’ ordering for theories allows us to readdress the question of a ‘better than’ ordering for explanations. Here the question is this. How does the probability that the theory provides an explanation trade off against the quality of the theory in evaluating the quality of the explanation? Again, I suggest a lexicographic criterion with the quality of the theory dominating.

Thus:

(18) If $\langle K, I, A \rangle$ and $\langle K', I, A' \rangle$ are explanations for $X$ then:

$\langle K, I, A \rangle$ better than $\langle K', I, A' \rangle$
	niff

$\langle K, I \rangle$ strictly better than $\langle K', I \rangle$

or

$\langle K, I \rangle$ equivalent $\langle K', I \rangle$
and

\[ P(X \text{ part of } A) > P(X \text{ part of } A'). \]

This 'better than' ordering for explanations clearly emphasizes the quality of the theory rather than the probability that the theory explains. Is this plausible? The main argument for its plausibility is this. The purpose of explaining \( X \) is to show that \( X \) can be integrated into our existing body of knowledge – without substantially revising that body. Explanation is reducing the \textit{prima facie} unfamiliar to the familiar. Among other things, this makes us more comfortable with the already familiar. We feel comfortable that existing knowledge is adequate – we need not take the trouble to learn something new. But, some things are more familiar than others. Integrating \( X \) into the framework of a more probable theory is just 'more satisfying'.

\textit{Department of Humanities and Social Sciences, Colorado School of Mines, Golden, CO 80401, U.S.A.}

\textbf{NOTE}

\footnote{I am indebted to \textbf{Mr.} Michael Pierce for calling my attention to these problems and discussing them with me.}

\textbf{REFERENCES}

I am sympathetic to Joe Sneed's general structural viewpoint as well as his Bayesian approach. The analysis of structure, the characterization of scientific structures, and the analysis of the nature of the structures, as well as a strong Bayesian bent, have been long-standing features of my own work in the philosophy of science. On the other hand, as the years go by, I find myself increasingly skeptical of very general philosophical accounts of explanation or causality or even of the viability of very general Bayesian methods. My view of science has moved increasingly from that of a foundationalist to the viewpoint that the conceptual content of science is best analyzed in terms of a diverse set of methods for solving a wide variety of problems. Schemes of great generality about scientific explanation, such as those of Sneed, can perhaps be useful in providing some kind of general framework, but I think they miss the main part of what we intuitively want from good scientific explanations. Moreover, our sense of satisfaction with good explanations is particularistic in nature and not in any natural way subsumed under the general schemes proposed by Joe. Above all, I am skeptical that a
general notion of part, as he uses it, can do the work he would like to have it do.

For those raised in traditional philosophy or even traditional philosophy of science, with the search for generality and universality of conceptual schemes so dominant, it is not easy to accept or even be sympathetic with a view that is skeptical of the success of any of the general schemes aimed at providing a traditional philosophy of science.

The picture of scientific activity I increasingly favor myself is closer to that of apprenticeship than to the propositional organization of knowledge. Perhaps only in mathematics do we have the generality of structure and generality of result that would satisfy the hearts as well as the minds of philosophers. Even physics, the most sophisticated of the empirical sciences, has only a pretense at great generality. Individual problems must be tackled by individual methods. The methods that are used to attack a particular problem depend upon the experience and insight of the investigator, not upon the sharp and exact codification of theory and its range of application. It is only a myth engendered by philosophers — even in the past to some extent by myself — that the deductive organization of physics in nice set-theoretical form is an achievable goal. A look at the chaos in the current literature in any part of physics is enough to quickly dispel that illusion. This does not mean that set-theoretical work cannot be done, it is just that its severe limitations must be recognized.

There are many ways of expanding upon the views I have just expressed. A central one is the current realization of how few even simple physical systems can be thoroughly understood, in the sense that detailed and precise predictions about the behavior of the system can be successfully verified. In a way, this is a lesson that was ready for understanding already in the nineteenth century in terms just alone of the massive and unsuccessful effort to master the three-body problem in classical mechanics. But the modern developments of chaos have made the facts much more salient and very much increased the awareness of how difficult it is to make successful predictions. But once the rarity of systems or structures whose behavior can be successfully predicted is recognized then much of the older talk about general schemes of explanation seems unsatisfactory for dealing with the rich details of actual science. This is not the place for a full-scale exposition of my views, but let me try to give one or two examples.
I want to contrast our attitude toward various cases of failure of prediction in classical mechanics. Let us begin with a case already mentioned, that of the three-body problem. Our inability to solve the differential equations, even in principle, in a satisfactory analytical form has been recognized for over 100 years. On the other hand, the derivation of the differential equations governing the motion of three bodies acted upon by the force of gravitation alone is an easy exercise. There is a very general belief that the differential equations accurately reflect many real situations to an extraordinarily high degree of accuracy. There is no need to say 'with complete accuracy' for no real systems are sufficiently isolated, just to mention one central reason. What is important is the belief that the equations hold to a very high degree of accuracy. But they are unmanageable from the standpoint of solutions.

A closely related example is that of a single body sliding or rolling down an inclined surface of variable degrees of roughness. We can write a differential equation including friction that also, because of the complicated nature of the surface, we cannot solve except numerically. But in this case the status of the equation is quite different. We do not think that it is possible to write down, to the same degree of precision at all, a differential equation governing the complicated physical phenomena that reflect the interaction between the surface and the moving body. Because our feelings of definiteness about these two cases seem so strong, a theory of structural explanation should provide a very compelling account of their difference. But to provide this requires entering into the details in a way that is not a feature of the current work.

Let us continue the same line of examples. A favorite one of many, also used by Sneed, is the way in which we accept the explanation of the motion of billiard balls on a table based upon the mechanical laws of collision. However, if we consider a somewhat more complicated billiard ball – I have in mind the kind studied by Ornstein et al., Sinai, and others – we enter an entirely new realm. The most striking theorems are that when the obstacles on our new and wonderfully different billiard table are convex in their shape then we cannot distinguish – no matter how many observations we take – between the motion of the billiard balls following the usual laws of mechanical collision, and the motion of the ball being described by a probabilistic Markov process. It is one of the great insights of modern mechanics, whose philosophical importance I have tried to stress in several publications over the years, that the separation between deterministic mechanical systems and ran-
dom probabilistic ones is not at all what it was once thought to be. Structuralist views of this kind of example are as yet missing from the literature.

The examples just cited are in a way too mathematical in character. The analysis of a physical problem we hope to solve, think we can solve and are willing to tackle, is more open-ended, less honed down to a few sets of variables than the kinds of examples I have just given. Here the apprenticeship of the physicist seems to me of the greatest importance. To know what to count and what to discard as unimportant in analyzing the given physical situation is not something wherein no end of training in mathematics and in the solution of differential equations will be much of help. As we move from applications of theory to a new situation and the organization of experiments, what I have to say is even more true. Moreover, in complicated high-technology experiments involving large numbers of individuals it is fair to say that no one commands any longer all the details of the experiments. It is not just apprenticeship, but a collection of mature apprenticeships, that are required to organize, prepare and execute the experiments planned. Individual papers with more than 100 authors are now not uncommon in high-energy physics, yet the obviously social nature of these experiments and the complicated sets of skills required to execute them have as yet received little attention in the analysis of physics by philosophers of science. In making these remarks about the collective enterprise of running experiments in high-energy physics, I am not interested as such in the sociology of science but more in the bewildering variety of patches of theory and patches of skills required to put the whole thing together. It is that it would be wonderful to have a detailed structural analysis of. Perhaps Sneed and his energetic collaborators can be persuaded to enlighten us all on some of these detailed matters.