Explanation as Unification

Article *in* Synthese · July 1999

DOI: 10.1023/A:1005214721929

CITATIONS
54

READS
137

1 author:

Gerhard Schurz
Heinrich-Heine-Universität Düsseldorf
171 PUBLICATIONS **1,989** CITATIONS

SEE PROFILE

GERHARD SCHURZ

EXPLANATION AS UNIFICATION

1. INTRODUCTION

We start from a *pragmatic* and *dynamic* modelling of explanations as *question–answer-episodes* (this is common to many recent approaches, cf. Bromberger 1965; van Fraassen 1980, Tuomela 1981; Schurz 1983; Stegmüller 1983; Sintonen 1984). Here an explanation episode consists of (at least) four elements: the explanation-seeking question ?P, the cognitive state C of the questioner where P is *in need of explanation*, the answer A, and the cognitive state C + A after receiving the answer, where C + A arises from C by expanding or revising C with A (in the sense of Gärdenfors 1988). If the answer A is indeed *explanatory* (with respect to P in C), then P is *explained* or *understood* in C + A, the need for explanation of P is *satisfied* and, thus, more or less extinguished in C + A.

There are different kinds of explanations which become important towards the end of this paper (Section 6). We first concentrate on (non-purposive) *why*-explanations, which were the focus of the debate on scientific explanation. Here the question has the standard form *Why P?* and the answer the standard form *P because of the reasons* (or premises) *Prem*.[1] Hence, the answer – if it is *complete* and not *elliptical* – makes two claims, namely that the statements in Prem are true and that the inference from Prem to P, abbreviated as Prem $\Rightarrow$ P, is correct in a *broad* sense (i.b.s.). We admit as *correct* (i.b.s.) only those inferences which establish an at least partial *information-transfer*, a partial reduction, or as we shall prefer to say, a partial *assimilation* of P to Prem. Thus, only deductive or probabilistic inferences and certain variations thereof will count as correct (i.b.s.). Moreover, we distinguish between *two levels* of explanation: (a) the explanation of *singular events* (where P is a singular proposition and Prem includes general propositions) and (b) the explanation of *laws* (where P is a general proposition and Prem includes elements of higher level theories).

When is the answer A *explanatory*? Two major *paradigms* have been intensively explored in the explanation debate:

Synthese **120:** 95–114, 1999.

(1) The *nomic expectability* approach, where the answer A explains why P iff it makes P predictable or expectable (Hempel 1965; Tuomela 1981; Niiniluoto 1981) or at least *increases* P's expectability (Gärdenfors 1980; Stegmüller 1983; van Fraassen 1980).

(2) The *causality* approach, where the answer A explains why P iff A gives a complete list of P's causes or causally relevant factors (Hempel 1977; Salmon 1984; Humphreys 1981; Lenk 1985, etc.).

The major shortcoming of paradigm (1) has been extensively discussed and follows from the fact that in explanation-seeking why-questions we ask for *reasons for being* – or causes – of P, but not merely for *reasons for believing* P. This distinction (which was introduced by Stegmüller 1969) is not at all dichotomic; in many cases, reasons for being are also reasons for believing and vice versa, but *not always*. This has been shown by a multitude of examples (Grünbaum 1963; Scriven 1962; Hempel 1965; Schurz 1983). For example, the red shift in the spectrum of remote stars (R) and the expansion of the universe (E) are biconditionally related, but only the inference from E to R is an explanation, while that from R to E gives a mere reason for believing.

The advantage of paradigm (1) is that expectability accounts are applicable not only to the explanation of singular events, but also to the explanation of laws, which are the scientifically *more* important kind of explanation (Alston 1971). In contrast, causality approaches are – at least prima facie – only applicable to explanations of singular events, because *causality* is defined by a theory of *causal processes* among spatiotemporally located events. As a further disadvantage of paradigm (2), a *purely* causal approach leads to the effect that A can explain P even if A lowers P's probability or degree of expectability (Salmon 1984; Jeffrey 1971), which is an odd consequence to be discussed later. Summarized, what is *dissatisfactory* if we exclusively deal with paradigms (1) and (2) is that even in the *limited* and well demarcated domain of scientific why-explanations we cannot explicate all of them by one common approach, but need two mutually incoherent paradigms to cover them.

This paper is about a third and less exhausted paradigm: (3) *explanation as unification*. I want to show that this paradigm provides a more general, an indeed unified view on explanations. The idea that *unification* or *coherence* is the main goal of science has prominent defenders in philosophy and physics (Mach 1883; Whewell 1847; Feigl 1970). Yet there have not been many detailed approaches to unification or coherence (cf. Lehrer 1974; Friedman 1974; Kitcher 1981; Thagard 1992; Schurz 1988; Schurz/Lambert 1994; Bartelborth 1996), probably because these notions are even more complicated than nomic expectability or causation. Note

also that some of these authors (Lehrer and Thagard) use unification or coherence for the different purpose of explicating *confirmation* rather than explanation.

Prima facie unification as well as coherence are global features of sets of statements or phenomena. In contrast, both paradigms (1) and (2) conceive explanation as a *local* relation between the phenomenon P (the *explanandum*) and the explaining premises Prem (the *explanans*). So, in order to motivate the unification approach we have to answer two main questions: (1) *What has explanation to do with unification?*, and (2) *How can unification be adequately defined?* In Sections 2–3 we answer the first question, in Section 4 the second question, and Sections 5–6 will present some applications.

2. REASONS FOR UNIFICATION AS A KEY TO EXPLANATION

There are several reasons for unification as an essential ingredient of explanation. I think the most direct reason is the following. Recall the pragmatic question-answer-model of explanation. One of its central concepts is the being-in-need-of-explanation of P in C, a need which is satisfied in C + A. This can be the case only if the following condition is satisfied (Stegmüller 1969, 770; Scriven 1959; Kutschera 1972, 331; and Käsbauer 1976, 270):

(U) The explanatory premises Prem must be *less in need of explanation* (in C + A) than the explanandum P (in C).

For instance, "Peter is flying past the window in the third floor, because one second ago he was flying past the window in the fifth floor" is not only a predictively but even a causally adequate argument. Still it is *not* adequate as an explanation, because the cause is here just as much in need of explanation as the effect. This example shows that (U) is a missing necessary condition even in the area of the causal explanation of macrophysical events. (An example of the same sort is Lehrer's owl-mouse-example, 1974, 166). Condition (U) is also necessary for the explanation of laws. An example are Bohr's stability postulates for electrons in an atom in the year 1913. They successfully predicted the hydrogen spectrum, but they were incoherent with classical mechanics and thus were themselves strongly in need of explanation, whence most scientists at that time did not view them as a satisfying explanation of the atomic behaviour.

We may justify condition (U) not only by examples but also in a general way, based on the widespread view that explanation and understanding are

semantically related in the sense that an explanation (of P) is satisfying iff it provides understanding (of P).[2] For it is reasonable to argue that one cannot *understand something* (P) by means of some other thing (Prem) which one *has not understood*.

In condition (U), *being-in-need-of-explanation* is the crucial concept which leads to a unification- or coherence-based approach of explanation. The being-in-need-of-explanation of a phenomenon P in cognitive state C comes in degrees, and it depends on how well P fits into C or coheres with C. The total coherence of a cognitive state C is a function of the coherence of its parts. If a phenomenon P is assimilated to the premises Prem in C + A, then the decoherence of P in C is removed in C + A, but at the same time, new information units in Prem have been added to C which may itself produce new decoherence with respect to other parts of C+A. However, if condition (U) is satisfied, than the loss of coherence due to the addition of Prem to C must be smaller than the gain of coherence due to the assimilation of P to Prem in C + A (by means of the inference i.b.s. Prem $\Rightarrow$ P). Hence condition (U) implies that the answer can be explanatory only if the *total* coherence of the cognitive corpus has been increased because of its addition. This is how condition (U) leads to the explanation-as-unification approach.

The remaining question is why we prefer *unification* instead of *coherence*. This is quickly explained. In our account, coherence is afforded by inference relations i.b.s. between propositions expressing elementary phenomena. Thereby we must eliminate completely circular inference relations. But *coherence minus circularity* equals *unification*. For unification means to reduce as many phenomena as possible to as few principles plus basic phenomena as possible. This is why we prefer unification rather than coherence. Now assume we have a well-defined partial order relation $>_u$ between the cognitive states (of cognitive agents or systems) measuring their degree of unification. Then we can explicate our unification approach to explanations in a preliminary way as follows:

(E) A is an *explanatory satisfying answer* to the question Why-P? in the cognitive state C iff (i) A claims (for some Prem) Prem & Prem $\Rightarrow$ P, where Prem $\Rightarrow$ P is a premise-relevant correct inference i.b.s., and (ii) C + A $>_u$ C.

Definition (E) *combines a local* condition (i) with a global condition (ii). Condition (i) requires that the explanandum must be *locally* assimilated to Prem with the help of an inference i.b.s., moreover, the increase of *global unification* required in (ii) must be the *result* of that local assimilation, i.e., must be produced by those parts of A which assimilate P. This

is guaranteed by requiring *premise-relevance* for arguments i.b.s., which roughly speaking means that no premise is inferentially superfluous. With the help of this local-global-combination our account avoids one widespread objection to the unification account, namely that unification is a global matter while explanation is a local affair (cf. van Fraassen 1980, 109–10). In our approach, an explanation of P is a local affair with a global effect.

Condition (U) is not the only way to motivate the unification approach (E). Another way is based on the idea that to explain a phenomenon P means to *fit P into the background theory or system*, respectively. This idea was first suggested in Lambert (1998), elaborated in Schurz/Lambert (1994), and is also used by Bartelborth (1996). Also the concept of *fitting-into* combines a local with a global component: to fit something P into some larger system C means not only to give P a *local* place in C but often requires a rearrangement of other parts of C such that all parts fit nicely together again.

A possible objection to our account could be made by drawing a distinction between *correct* explanations versus *more or less good* explanations, and by arguing that global unification is not an appropriate condition for correct explanations, but only for more-or-less good explanations (Wilson 1985, 43ff). Indeed, since our concept of unification increase $>_u$ is a gradual one (coming in degrees), it fits much more to the concept of a more-or-less good explanation (the greater the unification increase, the better the explanation) than to the concept of (merely) correct explanation. However, I think any rigid borderline between merely correct and good explanations will involve much arbitrariness. Should one say, for instance, that the explanation of combustion by phlogiston is correct, but does not fit into our present background knowledge, or should one say that this explanation is wrong just *because* of that reason? Moreover, the scientifically much more *important* notion is that of a good, or better explanation. For, most accounts of theory evaluation are based on a variant of *inference to the best explanation*. The usual picture is that all of the competing theories (as bizarre as they might be) offer *some* explanations, but the question is which ones are the better or even best ones. For this question, account (E) seems to be just the right kind of approach.

How does the unification paradigm relate to the nomic expectability and the causality paradigm? Obviously, it is the global condition (ii) which is missing in traditional accounts to explanation. Nomic expectability models are usually characterized by conditions which are similar to our local requirement (i). Thus, in nomic expectability accounts, explanations will produce global unification increase only if the addition of Prem to C does

not produce new decoherence elsewhere. By the way, observe how the conflict between the high-probability-approaches (Hempel 1965; Tuomela 1981; Niiniluoto 1981) and the increase-of probability-approaches (van Fraassen 1980; Gärdenfors 1980) – which allow explanation even if the probability of P given Prem is low, as in the syphilis-paresis example – is resolved within our account: *increase of predictability* and *decrease of surprise* are just two kinds of local *probabilistic unification increase*.

3. ON THE RELATION BETWEEN CAUSALITY AND UNIFICATION

Since nomic expectability is insufficient for explanation, the major rival of the unification approach is not the expectability but the causality approach. What is the relation between unification and causality? Can the demand for *causal* reasons be solely explained in terms of unification increase – as it was claimed by Kitcher (1981), who by the way is much more reserved in later writings (e.g., 1989)? To say it briefly: I think that this is *not* possible in the area of explanations of singular events, but it is indeed possible for the explanations of general laws. Let me explain why.

Concerning the explanations of singular events, considering a biconditional law A $\leftrightarrow$ B where only the direction from A to B is the causal one. Why should it be more unifying to explain B-events by A-events than to explain A-events by B-events? A detailed analysis shows that if the law is deterministic, the situation is completely equivalent vis-á-vis unification, since every third event C must be correlated with A and with B to exactly the same probabilistic degree. Thus, in the area of explanations of singular events, the causal requirement does *not* follow from the unification requirement merely based on *inferences i.b.s.* Since we have *characterized* explanations (in a pre-systematic way) as arguments which provide *reasons for being* and not merely reasons for believing – a characterization that clearly fits our ordinary intuitions – it seems to follow that in the area of explanations of singular events, we have to *add* the causality requirement as an *extra* requirement, which goes beyond the idea of unification in a merely *inferential* (or information-theoretic) sense. We realize this requirement by requiring that if P describes a singular event, then only inferences i.b.s. Prem $\Rightarrow$ P which are *causally adequate* in C + A can increase the explanatory unification, i.e., the unification relevant for explanation. Thereby, an inference i.b.s. Prem $\Rightarrow$ P (with singular P) is called *causally adequate* in C if P's detailed theoretical description plus the *causality theory* CT in C imply or at least make it plausible that there exists a causal process which leads from the singular antecedens events mentioned in Prem to the event described by P.[3] Thereby, a causality theory

CT in C is a group of highest level theories of C which put *constraints* on the ways in which real events may influence each other. Examples will be given soon.

The important point is this: A deeper analysis of causality will quickly lead us *back to unification*, but now on the *level of theories*. For let us ask: what is the reason *why* we distinguish in our cognitive system between causal and noncausal regularities *at all*? The answer can only be: because we have a *causality theory* CT in C which implies that only *certain* regularities reflect directly a causal process, while others can only be indirect consequences of causal processes (e.g., effects of a common cause). Thus, we agree with van Fraassen (1980, 124), that causality is not a priori but *theory-relative*. CT will typically contain two groups of theoretical principles: (a) principles which concern the decomposition of macrophysical objects into smallest parts, and (b) principles describing the propagation of 'causal' forces, of fields, momentum and/or of energy, in space and time. These propagation laws are what we conceive as *causal mechanisms*.

But now – what *distinguishes* theoretical laws which we believe as describing the real causal mechanisms as opposed to theoretical laws which we understand in a purely *instrumentalistic* sense? I think, the *only* scientifically acceptable answer can be this: the theoretical laws which we *trust* to describe the real causal mechanisms are those which *unify* all know empirical regularities in a *superior* way. And only *because* these theoretical laws have this overall unification power, our belief in real causal processes as opposed to noncausal correlations is rationally justified. Summarized, *it is a result of the search for unification at the theoretical level that we introduce the distinction between reasons of being and reasons of believing at the singular event level*. It is a result of this search for theoretical unification that at the level of singular events not every assimilation by a covering law regularity will count as an explanation.

Let me illustrate this point by two scenarios of how science, and in particular physics, may evolve in the future. The *first scenario* corresponds to the causality theory which is probably shared by the majority of scientists in our day. Here it is assumed, (a) that matter is composed into atoms and subatomic particles as described by chemistry and particle physics, and (b) that all causal influences are reducible, at least in principle, to the propagation of momentum and/or energy forward in time with a finite velocity bounded by the velocity of light. This causality theory puts constraints on causal relations proclaimed by 'higher' sciences (like biology, psychology etc.). For instance, there cannot be a backwards causation (from the purpose to its means, like in some kinds of teleological world views), or there cannot exist a far distance interaction as claimed

in parapsychology (through thought transmission, etc.). Also, quantum mechanical far distance correlations cannot be causal in this scenario and must have some (still unknown) common cause. Indeed, most theories of causal explanation (for instance that of Salmon 1984) assume a variant of the causality theory of this scenario.

However, several contemporary scientists *reject* this causality theory. Based on quantum theory, they believe in far distance interaction, or based on quantum electro-dynamics, they believe in backwards causation, etc. Let us assume, in our *second scenario*, that physics develops into a direction where the previous picture of causality gets completely destroyed. Nothing is left from *local* causality – all one can say in this new scenario is that quantum systems are *globally* interacting systems, and causality is a global feature of these systems which cannot be distributed among its spatiotemporal parts. In this scenario, every *correlation* which is predicted by quantum theory will *reflect* an effect of global causality. Causality will no longer be an asymmetrical, but rather be a symmetrical relation of interaction (be it forward, backward or simultaneous in time). Hence the causality theory of this scenario will ultimately collapse into *Humean causality*: causality will be nothing more than correlation. Of course, scientists will still keep the 'causal terminology', because to some extent the idea of causation is an *inborn* idea (cf. Sperber 1995). However, their idea of 'global' causality is in effect tantamount with the fact that here causality collapses into mere correlation. In the cognitive system underlying this scenario, the distinction between reasons for being and reasons for believing will *no longer* make sense – at least for quantum systems. And the 'symmetry' between explanations and predictions, which was proclaimed by Hempel and then refuted by so many philosophers, will again hold in the cognitive system underlying this scenario.

What these two examples are intended to demonstrate is the *generality* of our unification account. It makes the causal requirement dependent on the causality theories of C which in turn are justified by their unification power. Our account is able to deal with the cognitive states of both scenarios: for both of them it will define a reasonable concept of explanation. In our approach, causality is *not* a priori concept – we even admit the *empty* causality theory, the Humean theory where causality collapses into mere correlation. On the other hand, causal approaches of Salmon's sort are only suitable for the cognitive state underlying the first scenario. Hence, causal approaches of this sort can be viewed as a *specialization* of our approach where one assumes the causality theory to be *fixed* in some (a priori) way.

It is *crucial* for our approach that *unification power* just is *the* major criterion for the realistic interpretability of high level theories – it is *the*

rational justification of the interpretation of the high level theories CT of C as describing the fundamental causal processes of nature. Various authors have doubted that *unification* always goes hand in hand with closeness to the truth in a realistic sense (cf. Salmon 1984, 260; Morrison 1990; Humphreys 1993). We agree that their objections apply to *some* accounts of unification, but not to our approach. By means of two constructions our account ensures that unification success and success in realistic truth approximation will go hand in hand.

(1) Of course, the unification of a set of statements KNOW which *represents* our knowledge may involve various *artificial* unification effects which have nothing to do with reality but only with aspects of its representation (cf. Humphreys 1993). In order to ensure that the degree of unification of KNOW reflects the degree of unification of *real* phenomena we decompose KNOW into 'minimal parts' such that every 'part' corresponds to one elementary phenomenon. This decomposition is afforded by the logical method of *relevant consequence elements* which is similar to the method of clauses in computational logic and explained elsewhere (Schurz 1991; Schurz/Lambert 1994). The result of this decomposition is the system K of relevant knowledge elements. The important effect of this decomposition is that it eliminates *spurious unifications* due to logical irrelevance or redundancy – for example, the famous conjunction paradox besetting Kitcher's unification approach (1981, 526), as well as various further logical paradoxes of the debate on deductive explanation (cf. Schurz/Lambert 1994, 3.1). Note that we do not claim that the relevant elements of K directly represent the elementary constituents of reality – we only claim that based on the knowledge given in KNOW, the relevant elements of K are *best representational approximations* of the elementary constituents of the *known part* of reality. The relevant decomposition of KNOW into elements depends on KNOW. If KNOW changes into KNOW* then what has figured as a relevant element in K may cease to figure as a relevant element in K*. In other words, also our views about what counts as an elementary phenomenon are not a priori but knowledge-dependent. But for a *given* KNOW, K is uniquely determined by the method of relevant elements.

(2) We must also ensure that unification cannot increase on the *cost* of empirical confirmation – rather, it should yield empirical confirmation as a *by-product*. For, what distinguishes *unification in science* from unification in religion or mysticism that science does not seek to unify just any kinds of fictitious or speculative phenomena. In the end, it seeks to unify the actually observed phenomena, called *data*.

Unification of hypotheses is only of value if they contribute to data unification. In our account, this second condition is satisfied by *three subconstructions*: (2.1) Unification is not based on just any kind of analogies etc., but solely on *information-theoretic* (i.e., deductive or probabilistic) inference relations i.b.s., so that any unification has to do with probabilities of truth transfer. (2.2) In our comparative unification approach, which works with unificatory *costs* and *gains*, only the *data* in K have an *intrinsic gain*. Hypotheses do not have any intrinsic gain, only an intrinsic cost; they may produce an *extrinsic* gain only by their unification effect. Finally (2.3), complete inferential circles are excluded. Altogether (2.1–3) avoid spurious unifications due to empirically contentless speculations (such as Kitcher's God-will-example in 1981).

4. A SKETCH OF THE COMPARATIVE CONCEPT OF UNIFICATION

In the following sketch we avoid all technical definitions and just refer the reader to Schurz/Lambert (1994) where they are stated. We merely explain the most important ingredients of unification and some aspects which are new (as compared to Schurz/Lambert 1994). We start with the *representation* of the *cognitive state C* of a given cognitive agent AG. C has two components: (a) the relevant elements of AG's *descriptive knowledge* KNOW of AG, denoted by K, and (b) the *inferential 'knowledge'* of AG, denoted by I. Thus, C = ⟨K, I⟩. Since every statement in K corresponds to one elementary phenomenon, we loosely identify the phenomena with their linguistic representations and thus speak of K as the set of *phenomena* known (or believed) by AG. I is the set of all inferences mastered by AG. These inferences, written in the general form Prem ⇒ Con (Prem for the set of premises and Con for the conclusion), may be *deductive* or *probabilistic* (with high or low conditional probability). As *rationality conditions on C* we require that all inferences in I are correct and that KNOW is closed under deductive and inductive high-probability inferences in I. Deductive correctness coincides with validity. An inductive inference is correct only if it satisfies the condition of *maximal specificity*, which roughly requires that the premises must contain all information known in C which is probabilistically relevant for the conclusion (cf. Hempel 1968; Fetzer 1981; Schurz 1988; 1995).

An explicit and complete answer to an explanation-seeking question ?P is formally a pair A = ⟨Prem, Prem ⇒ P⟩ (for it claims Prem to be true and Prem ⇒ Con to be correct i.b.s.; cf. Section 1). Prem is the *descriptive* and

Prem ⇒ P the *inferential* part of A. (An exception are explanations-about in Section 6.) C + A results from C = ⟨K, I⟩ by revising K with Prem, and by expanding I with Prem ⇒ P; thus C + A = ⟨K*Prem, I ∪ {Prem ⇒ P}⟩ (where K*Prem is the decomposition of KNOW*Prem into relevant elements, and * is a revision operator in the sense of Gärdenfors 1988). If the answer A is redundant, then some parts of A are already contained in C. But in order to have the effect of unification increase, at least some parts of A must be new with respect to C.

Cognitive states ⟨K, I⟩ are the states of *information systems*. The elements of K are the *information units* and the inferences in I *assimilate* information. We prefer to speak of *assimilation* instead of *reduction* because our theory of unification goes far beyond traditional ideas of reduction. We distinguish the phenomena in K according to their *assimilation status* as follows.

(1) *Actual (why-) assimilation*: Why-actual assimilation implies predictability or nomic expectability. Assume X is a subset of K.

  (1.1) *Strong assimilation*: A phenomenon P in K is strongly assimilated with respect to X if there exists a deductive or high probability inference Prem ⇒ P in I with Prem a subset of X. Theories of explanation in the tradition of Hempel (1965) are solely concerned with strong assimilation.

  (1.2) *Approximative assimilation*: A phenomenon P is approximately assimilated with respect to X if there exists a deductive inference Prem ⇒ P* in I with Prem a subset of X such that P* approximates P. Approaches to theory reduction are typically concerned with approximative assimilation (cf. Niiniluoto 1982; Stegmüller 1986, 246–253).

  An *actual assimilation* basis of K is any subset X of K such that every phenomenon in (K-X) is actually assimilated with respect to X. Traditionally, X is identified with the unification basis of K (cf. Friedman 1974). Our theory, however, introduces important distinctions.

(2) *Potential (how-possible) assimilation*: A phenomenon which is not actually assimilated may be assimilated at least in some heuristic or virtual way.

  (2.1) *Random assimilation*: Events which have low probability (with respect to C) are not 'puzzling' or 'surprising' if it is known in C that they are random events, without any further 'hidden cause', such as the decay of a radioactive nucleus. Jeffrey (1971) and Salmon (1984, 109) have convincingly argued that even arguments which lower the probability of P may provide understanding of P

if they are causally complete and thus establish P as 'a matter of accident'. We define a phenomenon Pa in K as *random assimilated* with respect to C = ⟨K, I⟩ if there exists a low probability inference Prem ⇒ Pa in I with Prem a subset of K such that the probability law p(Px/Fx) = r in Prem is causally complete in the sense that no causally relevant strengthening F*x of Fx changes the conditional probability of Px according to K's causal probability laws.[4] Thus we acknowledge probability-decreasing arguments as explanations, but *not* as why-actual explanations, but rather as a kind of how-possible-explanations – namely possible *by chance*. Thereby we reconcile a deep conflict between the expectability and the causality paradigm of explanation.

(2.2) *Heuristic assimilation*: In heuristic assimilations, phenomena are inferred from theories T in K with help of initial or boundary conditions Cd which are not believed to be true but are merely *plausible* with respect to K – or as we say, which are not *in conflict* with K. This kind of assimilation plays an important role in science (cf. Kitcher 1981, 512–5). For example, the success of Darwin's theory of evolution was not so much due to its actual capacity to provide explanations of the evolution process, but rather due to its *promise* to be able to produce such explanations *if only* the initial and boundary conditions would be known. The notion of being in conflict with K is defined as follows. A *singular* phenomenon P *stands in conflict* with K if it is not randomly assimilated and there exists a low probability inference Prem ⇒ P in I with Prem a subset of K. (By the way, I think that this is also the proper explication a "surprising" event.) A *general* phenomenon (a law) L *stands in conflict* with K if K contains a theory T which is much more general than L but contains an ad-hoc restriction such that without this ad-hoc restriction L would be logically inconsistent with T. Hence, L describes a phenomenon which is exceptional with respect to to a general theory T of K – an example is the EPR paradox in the context of special relativity theory. We define a phenomenon P as *heuristically assimilated* with respect to K if there exists a theory T in K and a set of initial or boundary conditions Cd none of which stands in conflict with K such that T ∪ Cd ⇒ P is in I.

(3) *Dissimilated phenomena*: Dissimilated phenomena are 'riddles' or *anomalies* – even no 'how possible' explanation can be found for them, they resist all attempts of assimilation into the accepted theory. The EPR paradox is again an example. We define a phenomenon P as

dissimilated with respect to K if it stands in conflict with K, and for every theory T in K and every set Cd of initial or boundary conditions such that $T \cup Cd \Rightarrow P$ is an inference in I, at least one statement in Cd is itself in conflict with K. Our concept of dissimilation (or decoherence) is important for explicating what it means to be *in need of explanation* – it comes close to what Bromberger (1965, 82–3) has called a "p(uzzle)-predicament".

(4) *Basic Phenomena*, finally, are those phenomena which are neither actually nor potentially assimilated nor dissimilated. In scientifically developed cognitive systems, the only basic phenomena will be fundamental theories, because every fact or empirical law will fall into the range of some theory and, hence, will be (at least heuristically) assimilated or dissimilated.

A *unification classification* of K is a partition of K into four (disjoint) subsets K_a, K_p K_b, K_d of actually or potentially assimilated, basic and dissimilated phenomena, respectively. *The* unification classification of K is that unification classification which yields the greatest unification of K according to the criteria developed below.

The fewer the phenomena in K_d or K_b and the more the phenomena in K_a or K_p, the greater the unification of K will be. But for the estimation of unification it is not enough just to "count" the phenomena. We assume that every phenomenon P is associated with a certain *intrinsic weight* w(P), which reflects its *cognitive complexity*. We wish to avoid all arbitrariness in our system of intrinsic weights. All we assume is a partial ordering $>$ among the weights of phenomena. As a minimal rule for w(P), we assume that the cognitive complexity of a general phenomenon is greater than that of a singular one, and that of a theoretical phenomenon is greater than that of an observational one.

If a phenomenon P is newly added to K_b, its weight has to be paid as an *intrinsic cost*: $-w(P)$ (i.e., the negative weight). If, on the other hand, P is assimilated, this cost is saved, either completely or partially, depending on the strength of the assimilation. If P is dissimilated, its cost increases significantly. $-w_a(P)$, $-w_p(P)$ and $-w_d(P)$ denote the costs of an actually assimilated, potentially assimilated, and dissimilated phenomenon P, respectively. Concerning the absolute values, we have $w_d(P) \gg w(P) \gg w_p(P) > w_a(P) > 0$; $w_a(P)$ is zero in the case of deductive assimilation, and small but nonzero in case of high probability or approximative assimilation. Only the *data* in K have an *intrinsic gain*. For reasons of simplicity, we identify the intrinsic gain of a datum D with its (positive) intrinsic weight, w(D). This implies that adding new and basic data to K which do not affect other parts of K will cost and gain nothing.

The unification of two cognitive corpuses C and C^* is compared by the *method of shift diagrams*. This method allows us to determine the unification relation between C and C + A by considering only those phenomena which *change* their unification status during the transition from $K = \langle K_d, K_b, K_p, K_a\rangle$ to $K^*Prem = \langle K^*_d, K^*_b, K^*_p, K^*_a\rangle$. We describe this transition as a sequence of one element shifts. There are three kinds of shifts: a phenomenon P can be *added* to a set K_x (x in {d, b, p ,a}), which is denoted by P_{+x} it can be *subtracted* from K_x, denoted by P^*_{-x} or it can *move* from K_x to K_y, denoted by $P_{x\to y}$. A move can be decomposed into a sequence of a subtraction and an addition: $P_{x\to y} = (P_{-x}, P_{+y})$. The cost or gain of a shift is called its unification value, in short its *u-value*. Costs are negative u-values, gains are positive ones, written unsigned. The total unification effect of a sequence of shifts $(s_1, \ldots, s_n)$ with corresponding u-value $\langle u\rangle := (u_1, \ldots, u_n)$ leading from C to C^* is calculated by a simple principle: if each negative u-value u_i in $\langle u\rangle$ is balanced by a positive u-value u_j in $\langle u\rangle$ with an absolute value greater or equal u_j (following from the above weight rules), then the unification has not decreased, and if in addition at least one of the positive u-values dominates its negative counterpart or has no negative counterpart, then the unification has increased (for a proof cf. Schurz/Lambert 1994, appendix).

5. A SIMPLE EXAMPLE

We demonstrate the unification theory first at hand of our simple example of Peter who is flying past the window in the third floor (Section 2). Suppose the first answer you receive to your question "Why P?" is A1: "Because one second ago, Peter was flying past the window in the fifth floor". Though you are now able to infer P from A1 and your background theory (even in a causally adequate way), A1 does not provide real understanding since it is itself at least as dissimilated as P. Suppose instead that your question had evoked the answer A2: "Because the fire brigade is testing a new jumping sheet at our building". There is nothing puzzling about firebrigades testing jumping sheets: though the event is not very likely, it has plausible 'how possible'-explanations and thus is heuristically assimilated. Hence, the answer A2 is completely satisfying. The cognitive development underlying these two explanation episodes is illustrated in the two shift diagrams of Figure 1. A1des and A2des denote the descriptive parts of the answers A1 and A2 (A1des = Peter was flying past the window one second ago; A2des = fire-brigades are testing a new jumping sheet at our building).

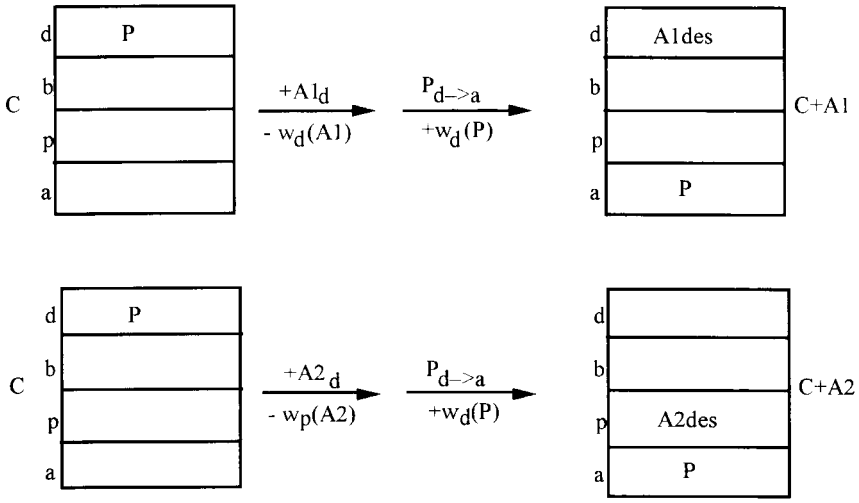

Figure 1. P = Peter is flying past the window in the third floor – Why? A1 = Because Peter was flying past the window in the fifth floor one second ago. A2 = Because fire-brigades are testing a new jumping sheet at our building.

We depict the unification classification of the knowledge system $\langle K_d, K_b, K_p, K_a \rangle$ graphically in a vertical notation: the more a phenomenon is assimilated, the nearer its distance to the bottom. This invokes the visual analogy with *minimalization of potential energy* – the less the 'potential energy' (cognitive cost) of the phenomena in C, the more 'stable' (unified) is C. In the case of A1, the u-gain $+w_d(P)$ due to assimilating P is balanced by the u-loss $-w_d(A1)$, hence there is no unification increase. In the case of A2, the u-gain $+w_d(P)$ dominates the u-loss $-w_p(A2)$ significantly, hence there is clear u-increase. The characteristic feature of the second answer is that its underlying cognitive development involves only additions (or moves) into K_a or K_p while no new basic or dissimilated phenomenon is obtained. We call such cognitive changes *consonant*, in distinction to *dissonant* changes as in the case of A1. Kuhn's 'normal' science is characterized by consonant u-increasing changes, while dissonant changes, if they grow too rapidly, lead to what Kuhn has called a 'revolutionary' stage.

Instead of presenting more examples we refer to Schurz/Lambert (1994) where the same kind of analysis is applied to an analysis of an *Aha!-experience* as consisting of a cascade of parallel strongly u-increasing moves (ch. 4.2), and to an analysis of the historical development of the atomic model (ch. 4.3).

6. KINDS OF EXPLANATIONS

We finally illustrate the generality of our approach by demonstrating how it covers *other* kinds of explanations beyond reason-giving why-explanations, and thus provides a unified account of explanation. I think it has too quickly been forgotten in the explanation debate that the concept of explanation in natural language is semantically rather *heterogeneous*. There are several *different* kinds of explanations. What they all have in common is that they *produce unification* (cf. Figure 2, the weights below the arrows are omitted; they can be retrieved from the shifted phenomena):

(1) In *why-actual-explanations*, the answer causes a move of P from K_d or K_b into K_a.
(2) In *how-possible explanations*, the explained phenomenon moves from K_d to K_p. The answer contains here also epistemic modal statements (like "R is a possible reason").
(3) In *how-explanations*, we know already something about the cause of a phenomenon P, and ask for more details about the causal process leading to P. Here the phenomenon moves from K_p to K_a.
(4) A further kind are *why-not-not-explanations*, which remove counteracting reasons of P: here P moves from K_d to K_b.
(5) *Explanation-about*: Here we don't understand a certain statement or phenomenon and just ask "explain P!". The answer will explain us P's *meaning* or P's *cognitive role* by demonstrating some important *consequences* of P. This kind of explanation has been completely neglected in the explanation debate, although it is easily accommodated in our approach: here, P is not the conclusion but an element of the *premise set* of the inference i.b.s. Prem $\Rightarrow$ Con, while Con is the descriptive part of the explanatory answer. By informing us about certain (so far unknown) consequences of P within C, the explanation-about-P produces unification increase. In science, explanations-about are especially important for the understanding of fundamental laws or theories. For example, if someone asks *why* bodies retain their velocity when no external forces act upon them, then Newtonian physics cannot provide a why-explanation of this phenomenon. But it can demonstrate how this principle allows one to infer many interesting phenomena. This improves our understanding of P and at the same time increases the unification of our cognitive state.
(6) Our analysis allows also a subtle treatment of *functional* or *teleological* explanations with the natural paraphrase *P is the case in order to produce effect E*. For instance, P may be the fact that birds have very light bones, and E the effect that this increases their ability to fly. Thus

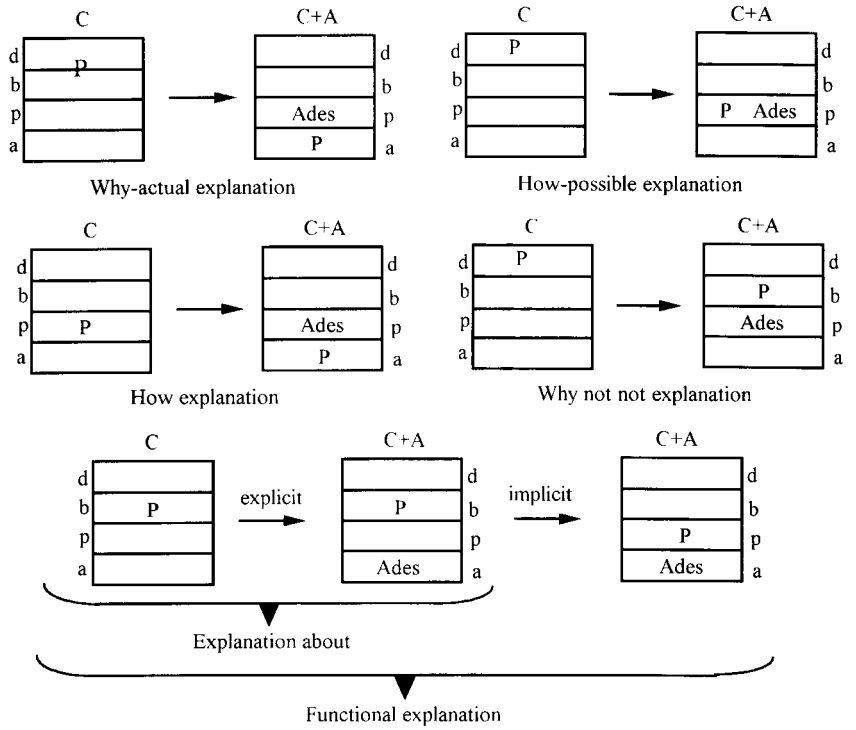

Figure 2. Kinds of Explanations: P = explained phenomenon, A = explanatory answer, Ades = descriptive part of A.

P is explained by pointing out that P has a certain effect E which has a high value V(E) for the underlying individual. Explicitly, functional explanations are explanations-about, because P is an element of the premise set, and the descriptive part of the answer is E. The scientific importance of a functional explanation, however, lies in this fact that at the same time it provides *implicitly* a *why-* or at least a *how-possible-*explanation. For there is a special background theory T in K, in our example *evolution theory*, which enables at least a heuristic assimilation of P whenever P has high value for the survival of the species to which the individual belongs. A similar case is the teleological explanation of actions. Here an answer like "Person p did action A in order to reach goal G" generates, by common sense background theory, the why-explanation that p did A because p intended to reach G.

Our analysis of functional explanations accords with the criticism of Hempel (1959) and Cummins (1975) against teleological explanations in the sense of Wright (1976). They argue that the functional effect E can never deductively explain the phenomenon P (which is a *means*

for E) because there exist functional equivalents. Thus explicitly, functional explanations are explanations-about (P causes E, and not vice versa). However, given particular circumstances C of the evolution of birds, we can indeed imagine an explanation of the evolution of light bones (P) where the fact that light bones increase the flying ability (E) figures as one element of the explanans premises which explain why bird-ancestors with light bones had a selective advantage in the sense of Neander (1991). These circumstances are usually unknown. So, what we can generate in this way is usually only a how-possible explanation, i.e., a heuristic assimilation.

NOTES

[1] Note that we abstract from the complication of a *contrast-class* "why P rather than Q_1, ..., Q_n". Contrast classes are investigated by Hansson (1974) and van Fraassen (1980). To abstract from a contrast class is tantamount to considering always the *minimal* contrast class "why P rather than not-P?". All aspects of our analysis can be generalized to explanation-seeking questions *with* contrast-classes.

[2] Scriven 1959; Bromberger 1965, 80; Hempel 1977, 100; Friedman 1974, 6; Salmon 1978; Kitcher 1981, 508; Tuomela 1980, 212; Achinstein 1983, 16; an exception is van Fraassen 1985, 642.

[3] Literally, this condition applies only to laws of temporal succession, though a similar condition can be set up also for laws of temporal coexistence (Schurz 1983; 1995).

[4] This is a much stronger condition than maximal specifity: the former excludes only statistically relevant strengthenings F^*a which are *known* about a.

REFERENCES

Achinstein, P.: 1983, *The Nature of Explanation*, Oxford University Press, Oxford.

Alston, W. P.: 1971, 'The Place of the Explanation of Particular Facts in Science', *Philosophy of Science* **38**, 13–34.

Bartelborth, T.: 1996, *Begründungsstrategien*, Akademie Verlag, Berlin.

Bromberger, S.: 1965, 'An Approach to Explanation', in R. Butler (ed.), *Analytical Philosophy, Second Series*, Basil Blackwell, Oxford, pp. 72–105.

Cummins, R.: 1975, 'Functional Analysis', *Journal of Philosophy* **72**, 741–65.

Feigl, H.: 1970, 'The Orthodox View of Theories: Remarks in Defense as well as Critique', in: *Minnesota Studies in the Philosophy of Science*, Vol. IV, University of Minnesota Press, Minneapolis.

Fetzer, J.: 1981, 'Probability and Explanation', *Synthese* **48**, 71-408.

Friedman, M.: 1974, 'Explanation and Scientific Understanding', *Journal of Philosophy* **71**, 5–19.

Gärdenfors, P: 1980, 'A Pragmatic Approach to Explanation', *Philosophy of Science* **47**, 404–423.

Gärdenfors, P.: 1988, *Knowledge in Flux*, MIT, Cambridge/Mass.
Grünbaum, A.: 1963, 'Temporally Asymmetry Principle', in Baumrin, B. (ed.), *Philosophy of Science. The Delaware Seminar*, Vol. I, J. Wiley, New York.
Hansson, B.: 1974, 'Explanations-of-What?', Mimeographed manuscript.
Hempel, C. G.: 1959, 'The Logic of Functional Analysis', reprinted in: Hempel (1965, 297–330).
Hempel, C. C.: 1965, *Aspects of Scientific Explanation and Other Essays*, Free Press, New York.
Hempel, C. G.: 1968, 'Maximal Specifity and Lawlikeness in Probabilistic Explanation', *Philosophy of Science* **35**, 116–133.
Hempel, C. C.: 1977, 'Nachwort 1976: Neuere Ideen zu den Problemen der statistischen Erklärung', in: C. G. Hempel, *Aspekte wissenschaftlicher Erklärung*, W. de Gruyter, Berlin, pp. 98–123.
Humphreys, P. : 1981, 'Aleatory Explanation', *Synthese* **48**, 225–232.
Humphreys, P.: 1993, 'Greater Unification Equals Greater Understanding?', *Analysis* **53**, 183–188.
Jeffrey, R. C.: 1971, 'Statistical Explanation vs. Statistical Relevance', in Salmon (1971), pp. 19–28.
Käsbauer, M.: 1976, 'Definitionen der wissenschaftlichen Erklärung', *Erkenntnis* **10**, 255–273.
Kitcher, P.: 1981, 'Explanatory Unification', *Philosophy of Science* **48**, 507–531.
Kitcher, P.: 1989, 'Explanatory Unification and the Causal Structure of the World', in Kitcher/Salmon (eds., 1989), pp. 410–505.
Kitcher, P. and Salmon, W. (eds.): 1989, *Scientific Explanation*, Minnesota Studies in the Philosophy of Science, Vol. XIII, University of Minnesota Press, Minneapolis.
Kutschera, F. v.: 1972, *Wissenschaftstheorie*, Vol. I and II, W. Fink, Munich.
Lehrer, K.: 1974, *Knowledge*, Clarendon Press, Oxford.
Lenk, H.: 1985, 'Bemerkungen zur pragmatisch-epistemischen Wende in der wissenschaftstheoretischen Analyse der Ereigniserklärungen', *Erkenntnis* **22**, 461–473.
Mach, E.: 1883 [1973], *Die Mechanik*, Wissenschatliche Buchgesellschaft, Darmstadt.
Morrison, M.: 1990, 'Unification, Realism and Inference', *British Journal for the Philosophy of Science* **41**, 305–332.
Neander, K.: 1991, 'Functions as Selected Effects: The Conceptual Analyst's Defense', *Philosophy of Science* **58**, 168–184.
Niiniluoto, I.: 1981, 'Statistical Explanation Reconsidered', *Synthese* **48**, 437–472.
Niiniluoto, I.: 1982, 'Truthlikeness for Quantitative Statements', *PSA* **I**, 208–216.
Salmon, W.: 1971, *Statistical Explanation and Statistical Relevance*, University of Pittsburgh Press, Pittsburgh.
Salmon, W.: 1978, 'Why ask 'Why?'?', *Proc. Adr. Amer. Phil. Assoc.* **51**, 683–705.
Salmon, W.: 1984, *Scientific Explanation and the Causal Structure of the World*, Princeton University Press, Princeton.
Schurz, G.: 1983, *Wissenschaftliche Erklärung*, dbv-Verlag der TU Graz, Graz.
Schurz, G.: 1988, 'Was ist wissenschaftliches Verstehen?', in Schurz (ed., 1988), pp. 235–298.
Schurz, G. (ed.): 1988, *Erklären und Verstehen in der Wissenschaft*, R. Oldenbourg (Scientia Nova), Munich.
Schurz, G.: 1991, 'Relevant Deduction', *Erkenntnis* **35**, 391–437.
Schurz, G.: 1995, 'Scientific Explanation: A Critical Survey', *Foundation of Science* **I**, 429–465.

Schurz, G./Lambert, K.: 1994, 'Outline of a Theory of Scientific Understanding', *Synthese* **101**, 65–120.
Scriven, M.: 1959, 'Truisms as the Grounds for Historical Explanation', in P. Gardiner (ed.), *Theories of History*, New York, pp. 443–468.
Scriven, M.: 1962, 'Explanation, Prediction and Laws', in H. Feigl and O. Maxwell (eds.), *Minnesota Studies in the Philosophy of Science*, Vol. III, University of Minnesota Press.
Sintonen, M.:1984, *The Pragmatics of Explanation*, Acta Philosophica Fennica **37**, Helsinki.
Sperber, D. et al. (eds.): 1995, *Causal Cognition*, Clarendon Press, Oxford.
Stegmüller, W.: 1969, *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie. Band I*, Wissenschaftliche Erklärung und Begründung, Springer, Berlin.
Stegmüller, W.: 1983, *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie. Band I: Erklärung – Begründung – Kausalität*. Zweite verbesserte und erweiterte Auflage, Springer, Berlin.
Stegmüller, W.: 1986, *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie. Band II. Dritter Teilband: Die Entwicklung des neuen Strukturalismus seit 1973*, Springer, Berlin.
Thagard, P.: 1992, *Conceptual Revolutions*, Princeton University Press.
Tuomela, R.: 1980. 'Explaining Explaining', *Erkenntnis* **15**, 211–243.
Tuomela, R.: 1981, 'Inductive Explanation', *Synthese* **48**, 257–294.
Van Fraassen, B.: 1980, *The Scientific Image*, Clarendon Press, Oxford.
Van Fraassen, B.: 1985, 'Salmon on Explanation', *Journal of Philosophy* **11**, 639–651.
Whewell, W.: 1847, *The Philosophy of the Inductive Sciences*, 2nd edition, 2 Volumes, John W. Parker, London.
Wilson, F.: 1985, *Explanation, Causation, and Deduction*, Reidel, Dordrecht.
Wright, L.: 1976, *Teleological Explanations*, University of California Press, Berkeley.

Department of Philosophy
Logic and Philosophy of Science Section
University of Salzburg
Franziskanergasse 1
Salzburg A5020
Austria