C

CARNAP, RUDOLF

(18 May 1891–14 September 1970)

Rudolf Carnap (1891–1970), preeminent member of the Vienna Circle, was one of the most influential figures of twentieth-century philosophy of science and analytic philosophy (including the philosophies of language, logic, and mathematics). The Vienna Circle was responsible for promulgating a set of doctrines (initially in the 1920s) that came to be known as logical positivism or logical empiricism (see Logical Empiricism; Vienna Circle). This set of doctrines has provided the point of departure for most subsequent developments in the philosophy of science. Consequently Carnap must be regarded as one of the most important philosophers of science of the twentieth century. Nevertheless, his most lasting positive contributions were in the philosophy of logic and mathematics and the philosophy of language. Meanwhile, his systematic but ultimately unsuccessful attempt to construct an inductive logic has been equally influential, since its failure has convinced most philosophers that such a project must fail.

Carnap was born in 1891 in Ronsdorf, near Barmen, now incorporated into the city of Wuppertal, in Germany (Carnap 1963a). In early childhood he was educated at home by his mother, Anna Carnap (neé Dörpfeld), who had been a schoolteacher. From 1898, he attended the Gymnasium at Barmen, where the family moved after his father's death that year. In school, Carnap's chief interests were in mathematics and Latin. From 1910 to 1914 Carnap studied at the Universities of Jena and Freiburg, concentrating first on philosophy and mathematics and, later, on philosophy and physics. Among his teachers in Jena were Bruno Bauch, a prominent neo-Kantian, and Gottlob Frege, a founder of the modern theory of quantification in logic. Bauch impressed upon him the power of Kant's conception that the geometrical structure of space was determined by the form of pure intuition. Though Carnap was impressed by Frege's ongoing philosophical projects, Frege's real (and lasting) influence came only later through a study of his writings. Carnap's formal intellectual work was interrupted between 1914 and 1918 while he did military service during World War I. His political views had already been of a mildly socialist/pacifist

nature. The horrors of the war served to make them more explicit and more conscious, and to codify them somewhat more rigorously.

Space

After the war, Carnap returned to Jena to begin research. His contacts with Hans Reichenbach and others pursuing philosophy informed by current science began during this period (see Reichenbach, Hans). In 1919 he read Whitehead and Russell's *Principia Mathematica* and was deeply influenced by the clarity of thought that could apparently be achieved through symbolization (see Russell, Bertrand). He began the construction of a putative axiom system for a physical theory of space-time. The physicists—represented by Max Wien, head of the Institute of Physics at the University of Jena—were convinced that the project did not belong in physics. Meanwhile, Bauch was equally certain that it did not belong in philosophy. This incident was instrumental in convincing Carnap of the institutional difficulties faced in Germany of doing interdisciplinary work that bridged the chasm between philosophy and the natural sciences. It also probably helped generate the attitude that later led the logical empiricists to dismiss much of traditional philosophy, especially metaphysics. By this point in his intellectual development (the early 1920s) Carnap was already a committed empiricist who, nevertheless, accepted both the analyticity of logic and mathematics and the Frege-Russell thesis of logicism, which required that mathematics be formally constructed and derived from logic (see Analyticity).

Faced with this lack of enthusiasm for his original project in Jena, Carnap (1922) abandoned it to write a dissertation on the philosophical foundations of geometry, which was subsequently published as *Der Raum.* Most traditional commentators have regarded the dissertation as a fundamentally neo-Kantian work because it included a discussion of "intuitive space," determined by pure intuition, independent of all contingent experience, and distinct from both mathematical (or formal) space and physical space (see Friedman 1999). However, recent reinterpretations argue for a decisive influence of Husserl (Sarkar 2003). In contrast to Kant, Carnap restricted what could be grasped by pure intuition to some topological and metric properties of finite local regions of space. He identifies this intuitive space with an infinitesimal space and goes on to postulate that a global space may be constructed from it by iterative extension. In agreement with Helmholtz and Moritz Schlick (a physicist-turned-philosopher, and founder of the Vienna Circle) (see Schlick, Moritz), the geometry of physical space was regarded as an empirical matter. Carnap included a discussion of the role of non-Euclidean geometry in Einstein's theory of general relativity. By distinguishing among intuitive, mathematical, and physical spaces, Carnap attempted to resolve the apparent differences among philosophers, mathematicians, and physicists by assigning the disputing camps to different discursive domains. In retrospect, this move heralded what later became the most salient features of Carnap's philosophical work: tolerance for diverse points of view (so long as they met stringent criteria of clarity and rigor) and an assignment of these viewpoints to different realms, the choice between which is to be resolved not by philosophically substantive (e.g., epistemological) criteria but by pragmatic ones (see Conventionalism).

The Constructionist Phase

During the winter of 1921, Carnap read Russell's *Our Knowledge of the External World.* Between 1922 and 1925, this work led him (Carnap 1963a) to begin the analysis that culminated in *Der logische Aufbau der Welt* ([1928] 1967), which is usually regarded as Carnap's first major work. The purpose of the *Aufbau* was to construct the everyday world from a phenomenalist basis (see Phenomenalism). The phenomenalist basis is an epistemological choice (§§54, 58). Carnap distinguished between four domains of objects: autopsychological, physical, heteropsychological, and cultural (§58). The first of these consists of objects of an individual's own psychology; the second, of physical entities (Carnap does not distinguish between everyday material objects and the abstract entities of theoretical physics); the third consists of the objects of some other individual's psychology; and the fourth, of cultural objects (*geistige Gegenstände*), which include historical and sociological phenomena.

From Carnap's ([1928] 1967) point of view, "[a]n object ... is called *epistemically primary* relative to another one ... if the second one is recognized through the mediation of the first and thus presupposes, for its recognition, the recognition of the first" (§54). Autopsychological objects are epistemically primary relative to the others in this sense. Moreover, physical objects are epistemically primary to heteropsychological ones because the latter can be recognized only through the mediation of the former—an expression on a face, a reading in an instrument, etc. Finally, heteropsychological objects are epistemically primary relative to cultural ones for the same reason.

The main task of the *Aufbau* is construction, which Carnap ([1928] 1967) conceives of as the converse of what he regarded as reduction (which is far from what was then—or is now—conceived of as "reduction" in Anglophone philosophy) (see Reductionism):

> [*A*]*n object is 'reducible' to others . . . if all statements about it can be translated into statements which speak only about these other objects* . . . . By *constructing* a concept from other concepts, we shall mean the indication of its "constructional definition" on the basis of other concepts. By a *constructional definition* of the concept *a* on the basis of the concepts *b* and *c*, we mean a rule of translation which gives a general indication how any propositional function in which *a* occurs may be transformed into a coextensive propositional function in which *a* no longer occurs, but only *b* and *c*. If a concept is reducible to others, then it must indeed be possible to construct it from them (§35).

AU:1

However, construction and reduction present different formal problems because, except in some degenerate cases (such as explicit definition), the transformations in the two directions may not have any simple explicit relation to each other. The question of reducibility/constructibility is distinct from that of epistemic primacy. In an important innovation in an empiricist context, Carnap argues that both the autopsychological and physical domains can be reduced to each other (in his sense). Thus, at the formal level, either could serve as the basis of the construction. It is epistemic primacy that dictates the choice of the former.

Carnap's task, ultimately, is to set up a constructional system that will allow the construction of the cultural domain from the autopsychological through the two intermediate domains. In the *Aufbau*, there are only informal discussions of how the last two stages of such a construction are to be executed; only the construction of the physical from the autopsychological is fully treated formally. As the basic units of the constructional system, Carnap chose what he calls "elementary experiences" (*Elementarerlebnisse,* or elex) (an extended discussion of Carnap's construction is to be found in Goodman 1951, chapter 5). These are supposed to be instantaneous cross-sections of the stream of experience—or at least bits of that stream in the smallest perceivable unit of time—that are incapable of further analysis. The only primitive relation that Carnap introduces is "recollection of similarity" (*Rs*). (In the formal development of the system, *Rs* is introduced first and the elex are defined as the field of *Rs.*) The asymmetry of *Rs* is eventually exploited by Carnap to introduce temporal ordering.

AU:2

Since the elex are elementary, they cannot be further analyzed to define what would be regarded as constituent qualities of them such as partial sensations or intensity components of a sensation. Had the elex not been elementary, Carnap could have used "proper analysis" to define such qualities by isolating the individuals into classes on the basis of having a certain (symmetric) relationship with each other. Carnap defines the process of "quasi-analysis" to be formally analogous to proper analysis but only defining "quasi-characteristics" or "quasi-constituents" because the elex are unanalyzable. Thus, if an elex is both *c* in color and *t* in temperature, *c* or *t* can be defined as classes of every elex having *c* or *t,* respectively. However, to say that *c* or *t* is a quality would imply that an elex is analyzable into simpler constituents. Quasi-analysis proceeds formally in this way (as if it is proper analysis) but defines only quasi-characteristics, thus allowing each elex to remain technically unanalyzable Quasi-analysis based on the relation "part similarity" (*Ps*), itself defined from *Rs*, is the central technique of the *Aufbau.* It is used eventually to define sense classes and, then, the visual sense, visual field places, the spatial order of the visual field, the order of colors and, eventually, sensations. Thus the physical domain is constructed out of the autopsychological. Carnap's accounts of the construction between the other two domains remain promissory sketches.

Carnap was aware that there were unresolved technical problems with his construction of the physical from the autopsychological, though he probably underestimated the seriousness of these problems. The systematic problems are that when a quality is defined as a class selected by quasi-analysis on the basis of a relation: (i) two (different) qualities that happen always to occur together (say, red and hot) will never be separated, and (ii) quality classes may emerge in which any two members bear some required relation to each other, but there may yet be no relation that holds between all members of the class. Carnap's response to these problems was extrasystematic: In the complicated construction of the world from the elex, he hoped that such examples would never or only very rarely arise. Nevertheless, because of these problems, and because the other constructions are not carried out, the attitude of the *Aufbau* is tentative and exploratory: The constructional system is presented as essentially unfinished. (Goodman 1951 also provides a lucid discussion of these problems.)

Some recent scholarship has questioned whether Carnap had any traditional epistemological

concerns in the *Aufbau.* In particular, Friedman (e.g., 1992) has championed the view that Carnap's concerns in that work are purely ontological: The *Aufbau* is not concerned with the question of the source or status of knowledge of the external world (see Empiricism); rather, it investigates the bases on which such a world may be constructed (see Richardson 1998). Both Friedman and Richardson—as well as Sauer (1985) and Haack (1977) long before them—emphasize the Kantian roots of the *Aufbau.* If this reinterpretation is correct, then what exactly the *Aufbau* owes to Russell (and traditional empiricism) becomes uncertain. However, as Putnam (1994, 281) also points out, this reinterpretation goes too far: Though the project of the *Aufbau* is not identical to that of Russell's external world program, there is sufficient congruence between the two projects for Carnap to have correctly believed that he was carrying out Russell's program. In particular, the formal constructions of the *Aufbau* are a necessary prerequisite for the development of the epistemology that Russell had in mind: One must be able to construct the world formally from a phenomenalist basis before one can suggest that this construction shows that the phenomena are the source of knowledge of the world. Moreover, this reinterpretation ignores the epistemological remarks scattered throughout the *Aufbau* itself, including Carnap's concern for the epistemic primacy of the basis he begins with. Savage (2003) has recently pointed out that the salient difference between Russell's and Carnap's project is that whereas the former chose sense data as his point of departure, the latter chose elementary experiences. But this difference is simply a result of Carnap's having accepted the results of Gestalt psychology as having definitively shown what may be taken as individual experiential bases; other than that, that is, with respect to the issue of empiricism, it has no philosophical significance.

In any case, by this time of his intellectual development, Carnap had fully endorsed not only the logicism of the *Principia*, but also the form that Whitehead and Russell had given to logic (that is, the ramified theory of types including the axioms of infinity and reducibility) in that work. However, Poincaré also emerges as a major influence during this period. Carnap did considerable work on the conceptual foundations of physics in the 1920s, and some of this work—in particular, his analysis of the relationship between causal determination and the structure of space—shows strong conventionalist attitudes (Carnap 1924; see also 1923 and 1926) (see Conventionalism; Poincaré, Henri).

Viennese Positivism

In 1926, at Schlick's invitation, Carnap moved to Vienna to become a *Privatdozent* (instructor) in philosophy at the University of Vienna for the next five years (see Vienna Circle). An early version of the *Aufbau* served as his *Habilitationsschrift.* He was welcomed into the Vienna Circle, a scientific philosophy discussion group organized by (and centered around) Schlick, who had occupied the chair for philosophy of the inductive sciences since 1922. In the meetings of the Vienna Circle, the typescript of the *Aufbau* was read and discussed. What Carnap seems to have found most congenial in the Circle—besides its members' concern for science and competence in modern logic—was their rejection of traditional metaphysics. Over the years, besides Carnap and Schlick, the Circle included Herbert Feigl, Kurt Gödel, Hans Hahn, Karl Menger, Otto Neurath, and Friedrich Waismann, though Gödel would later claim that he had little sympathy for the antimetaphysical position of the other members. The meetings of the Circle were characterized by open, intensely critical, discussion with no tolerance for ambiguity of formulation or lack of rigor in demonstration. The members of the Circle believed that philosophy was a collective enterprise in which progress could be made. These attitudes, even more than any canonical set of positions, characterized the philosophical movement—initially known as logical positivism and later as logical empiricism—that emerged from the work of the members of the Circle and a few others, especially Reichenbach. However, besides rejecting traditional metaphysics, most members of the Circle accepted logicism and a sharp distinction between analytic and synthetic truths. The analytic was identified with the *a priori*; the synthetic with the *a posteriori* (see Analyticity). A. J. Ayer, who attended some meetings of the Circle in 1933 (after Carnap had left—see below), returned to Britain and published *Language, Truth and Logic* (Ayer 1936) (see Ayer, Alfred Jules). This short book did much to popularize the views of the Vienna Circle among Anglophone philosophers, though it lacks the sophistication that is found in the writings of the members of the Circle, particularly Carnap.

Under Neurath's influence, during his Vienna years, Carnap abandoned the phenomenalist language he had preferred in the *Aufbau* and came to accept physicalism (see Neurath, Otto; Physicalism). The epistemically privileged language is one in which sentences reporting empirical knowledge of the world ("protocol sentences") employ terms

referring to material bodies and their observable properties (see Protocol Sentences). From Carnap's point of view, the chief advantage of a physicalist language is its intersubjectivity. Physicalism, moreover, came hand-in-hand with the thesis of the "unity of science," that is, that the different empirical sciences (including the social sciences) were merely different branches of a single unified science (see Unity of Science Movement). To defend this thesis, it had to be demonstrated that psychology could be based on a physicalist language In an important paper only published somewhat later, Carnap ([1932] 1934) attempted that demonstration (see Unity and Disunity of Science). Carnap's adoption of physicalism was final; he never went back to a phenomenalist language. However, what he meant by "physicalism" underwent radical transformations over the years. By the end of his life, it meant no more than the adoption of a nonsolipsistic language, that is, one in which intersubjective is possible (Carnap 1963b).

In the Vienna Circle, Wittgenstein's *Tractatus* was discussed in detail. Carnap found Wittgenstein's rejection of metaphysics concordant with the views he had developed independently. Partly because of Wittgenstein's influence on some members of the Circle (though not Carnap), the rejection of metaphysics took the form of an assertion that the sentences of metaphysics are meaningless in the sense of being devoid of cognitive content. Moreover, the decision whether a sentence is meaningful was to be made on the basis of the principle of verifiability, which claims that the meaning of a sentence is given by the conditions of its (potential) verification (see Verifiability). Observation terms are directly meaningful on this account (see Observation). Theoretical terms acquire meaning only through explicit definition from observation terms. Carnap's major innovation in these discussions within the Circle was to suggest that even the thesis of realism—asserting the "reality" of the external world—is meaningless, a position not shared by Schlick, Neurath, or Reichenbach. Problems generated by meaningless questions became the celebrated "pseudo-problems" of philosophy (Carnap [1928] 1967).

Wittgenstein's principle of verifiability posed fairly obvious problems in any scientific context. No universal generalization can ever be verified. Perhaps independently, Karl Popper perceived the same problem (see Popper, Karl Raimund). This led him to replace the requirement of verifiability with that of falsifiability, though only as a criterion to demarcate science from metaphysics, and not as one to be also used to demarcate meaningful from meaningless claims. It is also unclear what the status of the principle itself is, that is, whether it is meaningful by its own criterion of meaningfulness. Carnap, as well as other members of the Vienna Circle including Hahn and Neurath, realized that a weaker criterion of meaningfulness was necessary. Thus began the program of the "liberalization of empiricism." There was no unanimity within the Vienna Circle on this point. The differences between the members are sometimes described as those between a conservative "right" wing, led by Schlick and Waismann, which rejected both the liberalization of empiricism and the epistemological antifoundationalism of the move to physicalism, and a radical "left" wing, led by Neurath and Carnap, which endorsed the opposite views. The "left" wing also emphasized fallibilism and pragmatics; Carnap went far enough along this line to suggest that empiricism itself was a proposal to be accepted on pragmatic grounds. This difference also reflected political attitudes insofar as Neurath and, to a lesser extent, Carnap viewed science as a tool for social reform.

The precise formulation of what came to be called the criterion of cognitive significance took three decades (see Hempel 1950; Carnap 1956 and 1961) (see Cognitive Significance). In an important pair of papers, *Testability and Meaning*, Carnap (1936–1937) replaced the requirement of verification with that of confirmation; at this stage, he made no attempt to quantify the latter. Individual terms replace sentences as the units of meaning. Universal generalizations are no longer problematic; though they cannot be conclusively verified, they can yet be confirmed. Moreover, in *Testability and Meaning*, theoretical terms no longer require explicit definition from observational ones in order to acquire meaning; the connection between the two may be indirect through a system of implicit definitions. Carnap also provides an important pioneering discussion of disposition predicates.

The Syntactic Phase

Meanwhile, in 1931, Carnap had moved to Prague, where he held the chair for natural philosophy at the German University until 1935, when, under the shadow of Hitler, he emigrated to the United States. Toward the end of his Vienna years, a subtle but important shift in Carnap's philosophical interests had taken place. This shift was from a predominant concern for the foundations of physics to that for the foundations of mathematics and logic, even though he remained emphatic that the latter were important only insofar as they were used in the empirical sciences, especially physics.

In Vienna and before, following Frege and Russell, Carnap espoused logicism in its conventional sense, that is, as the doctrine that held that the concepts of mathematics were definable from those of logic, and the theorems of mathematics were derivable from the principles of logic. In the aftermath of Gödel's (1931) incompleteness theorems, however, Carnap abandoned this type of logicism and opted instead for the requirement that the concepts of mathematics and logic always have their customary (that is, everyday) interpretation in all contexts. He also began to advocate a strong conventionalism regarding what constituted "logic."

Besides the philosophical significance of Gödel's results, what impressed Carnap most about that work was Gödel's arithmetization of syntax. Downplaying the distinction between an object language and its metalanguage, Carnap interpreted this procedure as enabling the representation of the syntax of a language within the language itself. At this point Carnap had not yet accepted the possibility of semantics, even though he was aware of some of Tarski's work and had had some contact with the Polish school of logic. In this context, the representation of the syntax of a language within itself suggested to Carnap that all properties of a language could be studied within itself through a study of syntax.

These positions were codified in Carnap's major work from this period, *The Logical Syntax of Language* (Carnap 1934b and 1937). The English translation includes material that had to be omitted from the German original due to a shortage of paper; the omitted material was separately published in German as papers (Carnap 1934a and 1935). Conventionalism about logic was incorporated into the well-known principle of tolerance:

> *It is not our business to set up prohibitions but to arrive at conventions* [about what constitutes a logic] . . . . *In logic, there are no morals.* Every one is at liberty to build up his own logic, i.e., his own form of language, as he wishes. All that is required is that, if he wishes to discuss it, he must state his method clearly, and give syntactic rules instead of philosophical arguments.
>
> (Carnap 1937, 51–52; emphasis in the original)

Logic, therefore, is nothing but the syntax of language.

In *Syntax*, the principle of tolerance allows Carnap to navigate the ongoing disputes between logicism, formalism, and intuitionism/constructivism in the foundations of mathematics without abandoning any insight of interest from these schools. Carnap begins with a detailed study of the construction of two languages, I and II. The last few sections of *Syntax* also present a few results regarding the syntax of any language and also discuss the philosophical ramifications of the syntactic point of view. (Sarkar 1992 attempts a comprehensible reconstruction of the notoriously difficult formalism of *Syntax*.)

Language I, which Carnap calls "definite," is intended as a neutral core of all logically interesting languages, neutral enough to satisfy the strictures of almost any intuitionist or constructivist. It permits the definition of primitive recursive arithmetic and has bounded quantification (for all x up to some upper bound) but not much more. Its syntax is fully constructed formally. Language II, which is "indefinite" for Carnap, is richer. It includes Language I and has sufficient resources for the formulation of all of classical mathematics, and is therefore nonconstructive. Moreover, Carnap permits descriptive predicates in each language. Thus, the resources of Language II are strong enough to permit, in principle, the formulation of classical physics. The important point is that because of the principle of tolerance, the choice between Languages I and II or, for that matter, any other syntactically specified language, is not based on factual considerations. If one wants to use mathematics to study physics in the customary way, Language II is preferable, since as yet, nonconstructive mathematics remains necessary for physics. But the adoption of Language II, dictated by the pragmatic concern for doing physics, does not make Language I incorrect. This was Carnap's response to the foundational disputes of mathematics: By tolerance they are defined out of existence.

The price paid if one adopts the principle of tolerance is a radical conventionalism about what constitutes logic. Conventionalism, already apparent in Carnap's admission of both a phenomenalist and a physicalist possible basis for construction in the *Aufbau*, and strongly present in the works on the foundations of physics in the 1920s, had now been extended in *Syntax* to logic. As a consequence, what might be considered to be the most important question in any mathematical or empirical context—the choice of language—became pragmatic. This trend of relegating troublesome questions to the realm of pragmatics almost by fiat, thereby excusing them from systematic philosophical exploration, became increasingly prevalent in Carnap's views as the years went on.

Syntax contained four technical innovations in logic that are of significance: (i) a definition of analyticity that, as was later shown by S. C. Kleene, mimicked Tarski's definition of truth for a formalized language; (ii) a proof, constructed by Carnap

independently of Tarski, that truth cannot be defined as a syntactic predicate in any consistent formalized language; (iii) a rule for infinite induction (in Language I) that later came to be called the omega rule; and (iv), most importantly, a generalization of Gödel's first incompleteness theorem that has come to be called the fixed-point lemma. With respect to (iv), what Carnap proved is that in a language strong enough to permit arithmetization, for any syntactic predicate, one can construct a sentence that would be interpreted as saying that it satisfies that predicate. If the chosen predicate is unprovability, one gets Gödel's result.

Besides the principle of tolerance, the main philosophical contribution of *Syntax* was the thesis that philosophy consisted of the study of logical syntax. Giving a new twist to the Vienna Circle's claim that metaphysical claims were meaningless, Carnap argues and tries to show by example that sentences making metaphysical claims are all syntactically ill-formed. Moreover, since the arithmetization procedure shows that all the syntactic rules of a language can be formulated within the language, even the rules that determine what sentences are meaningless can be constructed within the language. All that is left for philosophy is a study of the logic of science. But, as Carnap (1937) puts it: "The *logic of science* (logical methodology) is nothing else than the *syntax of the language of science*. . . . To share this view is to *substitute logical syntax for philosophy*" (7–8; emphasis in original). The claims of *Syntax* are far more grandiose—and more flamboyant—than anything in the *Aufbau.*

Semantics

In the late 1930s Carnap abandoned the narrow syntacticism of *Syntax* and, under the influence of Tarski and the Polish school of logic, came to accept semantics. With this move, Carnap's work enters its final mature phase. For the first time, he accepted that the concept of truth can be given more than pragmatic content. Thereupon, he turned to the systematization of semantics with characteristic vigor, especially after his immigration to the United States, where he taught at the University of Chicago from 1936 to 1952. In his contribution to the *International Encyclopedia of Unified Science* (Carnap 1939), on the foundations of logic and mathematics, the distinctions among syntactic, semantic, and pragmatic considerations regarding any language are first presented in their mature form.

Introduction to Semantics, which followed in 1942, develops semantics systematically. In *Syntax* Carnap had distinguished between two types of transformations on sentences: those involving "the method of derivation" or "*d*-method," and those involving the "method of consequence" or "*c*-method." Both of these were supposed to be syntactic, but there is a critical distinction between them. The former allows only a finite number of elementary steps. The latter places no such restriction and is, therefore, more "indefinite." Terms defined using the *d*-method ("*d*-terms") include "derivable," "demonstrable," "refutable," "resoluble," and "irresoluble"; the corresponding *c*-terms are "consequence," "analytic," "contradictory," "L-determinate," and "synthetic." After the conversion to semantics, Carnap proposed that the *c*-method essentially captured what semantics allowed; the *c*-terms referred to semantic concepts.

Thus semantics involves a kind of formalization, though one that is dependent on stronger inference rules than the syntactical ones. In this sense, as Church (1956, 65) has perceptively pointed out, Carnap—and Tarski—reduce semantics to formal rules, that is, syntax. Thus emerges the interpretation of deductive logic that has since become the textbook version, so commonly accepted that is has become unnecessary to refer to Carnap when one uses it. For Carnap, the semantic move has an important philosophical consequence: Philosophy is no longer to be replaced just by the syntax of the language of science; rather, it is to be replaced by the syntax and the semantics of the language of science.

Carnap's (1947) most original—and influential—work in semantics is *Meaning and Necessity*, where the basis for an intensional semantics was laid down. Largely following Frege, intensional concepts are distinguished from extensional ones. Semantical rules are introduced and the analytic/synthetic distinction is clarified by requiring that any definition of analyticity must satisfy the (meta-) criterion that analytic sentences follow from the semantical rules alone. By now Carnap had fully accepted that semantic concepts and methods are more fundamental than syntactic ones: The retreat from the flamboyance of *Syntax* was complete. The most important contribution of *Meaning and Necessity* was the reintroduction into logic, in the new intensional framework, of modal concepts that had been ignored since the pioneering work of Lewis (1918). In the concluding chapter of his book, Carnap introduced an operator for necessity, gave semantic rules for its use, and showed how other modal concepts such as possibility, impossibility, necessary implication, and necessary equivalence can be defined from this basis.

By this point, Carnap had begun to restrict his analyses to exactly constructed languages, implicitly

AU:3

abandoning even a distant hope that they would have any direct bearing on natural languages. The problem with the latter is that their ambiguities made them unsuited for the analysis of science, which, ultimately, remained the motivation of all of Carnap's work. Nevertheless, Carnap's distinction between the analytic and the synthetic came under considerable criticism from many, including Quine (1951), primarily on the basis of considerations about natural languages (see Analyticity; Quine, Willard Van Orman). Though philosophical fashion has largely followed Quine on this point, at least until recently, Carnap was never overly impressed by this criticism (Stein 1992). The analytic/synthetic distinction continued to be fundamental to his views, and, in a rejoinder to Quine, Carnap argued that nothing prevented empirical linguistics from exploring intensions and thereby discovering cases of synonymy and analyticity (Carnap 1955).

Carnap's (1950a) most systematic exposition of his final views on ontology is also from this period. A clear distinction is maintained between questions that are internal to a linguistic framework and questions that are external to it. The choice of a linguistic framework is to be based not on cognitive but on pragmatic considerations. The external question of "realism," which ostensibly refers to the "reality" of entities of a framework in some sense independent of it, rather than to their "reality" within it after the framework has been accepted, is rejected as noncognitive (see Scientific Realism). This appears to be an anti-"realist" position, but it is not in the sense that within a framework, Carnap is tolerant of the abstract entities that bother nominalists. The interesting question becomes the pragmatic one, that is, what frameworks are fruitful in which contexts, and Carnap's attitude toward the investigation of various alternative frameworks remains characteristically and consistently tolerant.

Carnap continued to explore questions about the nature of theoretical concepts and to search for a criterion of cognitive significance, preoccupations of the logical empiricists that date back to the Vienna Circle. Carnap (1956) published a detailed exposition of his final views regarding the relation between the theoretical and observational parts of a scientific language. This paper emphasizes the methodological and pragmatic aspects of theoretical concepts. It also contains his most subtle, though not his last, attempt to explicate the notion of the cognitive significance of a term and thus establish clearly the boundary between scientific and nonscientific discourse. However, the criterion he formulates makes theoretical terms significant only with respect to a class of terms, a theoretical language, an observation language, correspondence rules between them, and a theory. Relativization to a theory is critical to avoiding the problems that beset earlier attempts to find such a criterion. Carnap proves several theorems that are designed to show that the criterion does capture the distinction between scientific and nonscientific discourse. This criterion was criticized by Roozeboom (1960) and Kaplan (1975), but these criticisms depend on modifying Carnap's original proposal in important ways. According to Kaplan, Carnap accepted his criticism, though there is apparently no independent confirmation of that fact. However, Carnap (1961) did turn to a different formalism (Hilbert's ε-operator) in what has been interpreted as his last attempt to formulate such a criterion (Kaplan 1975), and this may indicate dissatisfaction with the 1956 attempt. If so, it remains unclear why: That attempt did manage to avoid the technical problems associated with the earlier attempts of the logical empiricists (see Cognitive Significance).

Probability and Inductive Logic

From 1941 onward Carnap also began a systematic attempt to analyze the concepts of probability and to formulate an adequate inductive logic (a logic of confirmation), a project that would occupy him for the rest of his life. Carnap viewed this work as an extension of the semantical methods that he had been developing for the last decade. This underscores an interesting pattern in Carnap's intellectual development. Until the late 1930s Carnap viewed syntactic categories only as nonpragmatically specifiable; questions of truth and confirmation were viewed as pragmatic. His conversion to semantics saw the recovery of truth from the pragmatic to the semantic realm. Now, confirmation followed truth down the same pathway.

In *Logical Foundations of Probability* (1950b), his first systematic analysis of probability, Carnap distinguished between two concepts of probability: "statistical probability," which was the relevant concept to be used in empirical contexts and generally estimated from the relative frequencies of events, and "logical probability," which was to be used in contexts such as the confirmation of scientific hypotheses by empirical data. Though the latter concept, usually called the "logical interpretation" of probability, went back to Keynes (1921), Carnap provides its first systematic explication (see Probability).

Logical probability is explicated from three different points of view (1950b, 164–8): (i) as a

conditional probability $c(h,e)$, which measures the degree of confirmation of a hypothesis h on the basis of evidence e (if $c(h,e) = r$, then r is determined by logical relations between h and e); (ii) as a rational degree of belief or fair betting quotient (if $c(h,e) = r$, then r is a fair bet on h if e correctly describes the total knowledge available to a bettor); and (iii) as the limit of relative frequencies in some cases. According to Carnap, the first of these, which specifies a confirmation function (''c-function''), is the concept that is most relevant to the problem of induction. In the formal development of the theory, probabilities are associated with sentences of a formalized language.

In *Foundations*, Carnap (1950b) believed that a unique measure $c(h,e)$ of the degree of confirmation can be found, and he even proposed one (*viz*., Laplace's rule of succession), though he could not prove its uniqueness satisfactorily. His general strategy was to augment the standard axioms of the probability calculus by a set of ''conventions on adequacy'' (285), which turned out to be equivalent to assumptions about the rationality of degrees of belief that had independently been proposed by both Ramsey and de Finetti (Shimony 1992). In a later work, *The Continuum of Inductive Methods*, using the conventions on adequacy and some plausible symmetry principles, Carnap (1952) managed to show that all acceptable c-functions could be parameterized by a single parameter, a real number, $\lambda \in [0,\infty]$. The trouble remained that there is no intuitively appealing *a priori* strategy to restrict λ to some preferably very small subset of $[0,\infty]$. At one point, Carnap even speculated that it would have to be fixed empirically. Unfortunately, some higher-order induction would then be required to justify the procedure for its estimation, and potentially, this leads to infinite regress (see Confirmation Theory; Inductive Logic).

Carnap spent 1952–54 at the Institute for Advanced Study at Princeton, New Jersey, where he continued to work on inductive logic, often in collaboration with John Kemeny. He also returned to the foundations of physics, apparently motivated by a desire to trace and explicate the relations between the physical concept of entropy and an abstract concept of entropy appropriate for inductive logic. His discussion with physicists proved to be disappointing and he did not publish his results. (These were edited and published by Abner Shimony [Carnap 1977] after Carnap's death.)

In 1954 Carnap moved to the University of California at Los Angeles to assume the chair that had become vacant with Reichenbach's death in 1953. There he continued to work primarily on inductive logic, often with several collaborators, over the next decade. There were significant modifications of his earlier attempts to formulate a systematic inductive logic (see Carnap and Jeffrey 1971 and Jeffrey 1980. An excellent introduction to this part of Carnap's work on inductive logic is Hilpinen 1975). Obviously impressed by the earlier work of Ramsey and de Finetti, Carnap (1971a) returned to the second of his three 1950 explications of logical probability and emphasized the use of inductive logic in decision problems.

More importantly, Carnap, in *A Basic System of Inductive Logic* (1971b and 1980), finally recognized that attributing probabilities to sentences was too restrictive. If a conceptual system uses real numbers and real-valued functions, no language can express all possible cases using only sentences or classes of sentences. Because of this, he now began to attribute probabilities to events or propositions (which are taken to be synonymous). This finally brought some concordance between his formal methods and those of mathematical statisticians interested in epistemological questions. Propositions are identified with sets of models; however, the fields of the sets are defined using the atomic propositions of a formalized language. Thus, though probabilities are defined as measures of sets, they still remain relativized to a particular formalized language. Because of this, and because the languages considered remain relatively simple (mostly monadic predicate languages), much of this work remains similar to the earlier attempts.

By this point Carnap had abandoned the hope of finding a unique c-function. Instead, he distinguished between subjective and objective approaches in inductive logic. The former emphasizes individual freedom in the choice of necessary conventions; the latter emphasizes the existence of limitations. Though Carnap characteristically claimed to keep an open mind about these two approaches, his emphasis was on finding rational *a priori* principles that would systematically limit the choice of c-functions. Carnap was still working on this project when he died on September 14, 1970. He had not finished revising the last sections of the second part of the *Basic System*, both parts of which were published only posthumously.

Toward the end of his life, Carnap's concern for political and social justice had led him to become an active supporter of an African American civil rights organization in Los Angeles. According to Stegmüller (1972, lxvi), the ''last photograph we have of Carnap shows him in the office of this organization, in conversation with various members. He was the only white in the discussion group.''

The Legacy

Thirty-five years after Carnap's death it is easier to assess Carnap's legacy, and that of logical empiricism, than it was in the 1960s and 1970s, when a new generation of analytic philosophers and philosophers of science apparently felt that they had to reject that work altogether in order to be able to define their own philosophical agendas. This reaction can itself be taken as evidence of Carnap's seminal influence, but, nevertheless, it is fair to say that Carnap and logical empiricism fell into a period of neglect in the 1970s from which it began to emerge only in the late 1980s and early 1990s. Meanwhile it became commonplace among philosophers to assume that Carnap's projects had failed.

Diagnoses of this failure have varied. For some it was a result of the logical empiricists' alleged inability to produce a technically acceptable criterion for cognitive significance. For others, it was because of Quine's dicta against the concept of analyticity and the analytic/synthetic distinction (see Analyticity; Quine, Willard Van Orman). Some took Popper's work to have superseded that of Carnap and the logical empiricists (see Popper, Karl Raimund). Many viewed Kuhn's seminal work on scientific change to have shown that the project of inductive logic was misplaced; they, and others, generally regarded Carnap's attempt to explicate inductive logic to have been a failure (see Kuhn, Thomas; Scientific Change). Finally, a new school of "scientific realists" attempted to escape Carnap's arguments against external realism (see Scientific Realism).

There can be little doubt that Carnap's project of founding inductive logic has faltered. He never claimed that he had gone beyond preliminary explorations of possibilities, and, though there has been some work since, by and large, epistemologists of science have abandoned that project in favor of less restrictive formalisms, for instance, those associated with Bayesian or Neyman-Pearson statistics (see Bayesianism; Statistics, Philosophy of). But, with respect to every other case mentioned in the last paragraph, the situation is far less clear. It has already been noted that Carnap's final criterion for cognitive significance does not suffer from any technical difficulty no matter what its other demerits may be. Quine's dicta against analyticity no longer appear as persuasive as they once did (Stein 1992); Quine's preference for using natural—rather than formalized—language in the analysis of science has proved to be counterproductive; and his program of naturalizing epistemology has yet to live up to its initial promise. Putnam's "internal realism" is based on and revives Carnap's views on ontology, and Kuhn is perhaps now better regarded as having contributed significantly to the sociology rather than to the epistemology of science.

However, to note that some of the traditionally fashionable objections to Carnap and logical empiricism cannot be sustained does not show that that work deserves a positive assessment on its own. There still remains the question: What, exactly, did Carnap contribute? The answer turns out to be straightforward: The textbook picture of deductive logic that is in use today is the one that Carnap produced in the early 1940s after he came to acknowledge the possibility of semantics. The fixed-point lemma has turned out to be an important minor contribution to logic. The reintroduction of modal logic into philosophy opened up new vistas for Kripke and others in the 1950s and 1960s. Carnap's views on ontology continue to influence philosophers today. Moreover, even though the project of inductive logic seems unsalvageable to most philosophers, it is hard to deny that Carnap managed to clarify significantly the ways in which concepts of probability must be deployed in the empirical sciences and why the problem of inductive logic is so difficult. But, most of all, Carnap took philosophy to a new level of rigor and clarity, accompanied by an open-mindedness (codified in the principle of tolerance) that, unfortunately, is not widely shared in contemporary analytic philosophy.

SAHOTRA SARKAR

References

Ayer, A. J. (1936), *Language, Truth and Logic*. London: Gollancz.

Carnap, R. (1922), *Der Raum*. Berlin: von Reuther and Reichard.

——— (1923), "Über die Aufgabe der Physik und die Anwendung des Grundsatzes der Einfachsteit," *Kant-Studien* 28: 90–107.

——— (1924), "Dreidimensionalität des Raumes und Kausalität: Eine Untersuchung über den logischen Zusammenhang zweier Fiktionen," *Annalen der Philosophie und philosophischen Kritik* 4: 105–130.

——— (1926), *Physikalische Begriffsbildung*. Karlsruhe, Germany: Braun.

——— ([1928] 1967), *The Logical Structure of the World and Pseudoproblems in Philosophy*. Berkeley and Los Angeles: University of California Press.

——— ([1932] 1934. *The Unity of Science*. London: Kegan Paul, Trench, Trubner & Co.

——— (1934a), "Die Antinomien und die Unvollständigkeit der Mathematik," *Monatshefte für Mathematik und Physik* 41: 42–48.
——— (1934b), *Logische Syntax der Sprache*. Vienna: Springer.
——— (1935), "Ein Gültigskriterium für die Sätze der klassischen Mathematik," *Monatshefte für Mathematik und Physik* 42: 163–190.
——— (1936–1937), "Testability and Meaning," *Philosophy of Science* 3 and 4: 419–471 and 1–40.
——— (1937), *The Logical Syntax of Language*. London: Kegan Paul, Trench, Trubner & Co.
——— (1939), *Foundations of Logic and Mathematics*. Chicago: University of Chicago Press.
——— (1947), *Meaning and Necessity: A Study in Semantics and Modal Logic*. Chicago: University of Chicago Press.
——— (1950a), "Empiricism, Semantics, and Ontology," *Revue Internationale de Philosophie* 4: 20–40.
——— (1950b), *Logical Foundations of Probability*. Chicago: University of Chicago Press.
——— (1952), *The Continuum of Inductive Methods*. Chicago: University of Chicago Press.
——— (1955), "Meaning and Synonymy in Natural Languages," *Philosophical Studies* 7: 33–47.
——— (1956), "The Methodological Character of Theoretical Concepts," in H. Feigl and M. Scriven (eds.), *The Foundations of Science and the Concepts of Psychology and Psychoanalysis*. Minneapolis: University of Minnesota Press, 38–76.
——— (1961), "On the Use of Hilbert's ε-Operator in Scientific Theories," in Y. Bar-Hillel, E. I. J. Poznanski, M. O. Rabin, and A. Robinson (eds.), *Essays on the Foundations of Mathematics*. Jerusalem: Magnes Press, 156–164.
——— (1963a), "Intellectual Autobiography," in P. A. Schilpp (ed.), *The Philosophy of Rudolf Carnap*. La Salle, IL: Open Court, 3–84.
——— (1963b), "Replies and Systematic Expositions," in P. A. Schilpp (ed.), *The Philosophy of Rudolf Carnap*. La Salle, IL: Open Court, 859–1013.
——— (1971a), "Inductive Logic and Rational Decisions," in R. Carnap and R. C. Jeffrey (eds.), *Studies in Inductive Logic and Probability* (Vol. 1). Berkeley and Los Angeles: University of California Press, 5–31.
——— (1971b), "A Basic System of Inductive Logic, Part I," in R. Carnap and R. C. Jeffrey (eds.), *Studies in Inductive Logic and Probability* (Vol. 1). Berkeley and Los Angeles: University of California Press, 33–165.
——— (1977), *Two Essays on Entropy*. Berkeley and Los Angeles: University of California Press.
——— (1980), "A Basic System of Inductive Logic, Part II," in R. C. Jeffrey (ed.), *Studies in Inductive Logic and Probability* (Vol. 2). Berkeley and Los Angeles: University of California Press, 8–155.
Carnap, R., and R. C. Jeffrey (eds.) (1971), *Studies in Inductive Logic and Probability* (Vol. 1). Berkeley and Los Angeles: University of California Press.
Friedman, M. (1992), "Epistemology in the *Aufbau*," *Synthese* 93: 191–237.
——— (1999), *Reconsidering Logical Empiricism*. New York: Cambridge University Press.
Gödel, K. (1931), "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme 1," *Monatshefte für Mathematik und Physik* 38: 173–198.
Goodman, N. (1951), *The Structure of Experience*. Cambridge, MA: Harvard University Press.
Haack, S. (1977), "Carnap's *Aufbau*: Some Kantian Reflections," *Ratio* 19: 170–175.
Hempel, C. G. (1950), "Problems and Changes in the Empiricist Criterion of Meaning," *Revue Internationale de Philosophie* 11: 41–63.
Hilpinen, R. (1975), "Carnap's New System of Inductive Logic," in J. Hintikka (ed.), *Rudolf Carnap, Logical Empiricist: Materials and Perspectives*. Dordrecht, Netherlands: Reidel, 333–359.
Jeffrey, R. C. (ed.) (1980), *Studies in Inductive Logic and Probability* (Vol. 2). Berkeley and Los Angeles: University of California Press.
Kaplan, D. (1975), "Significance and Analyticity: A Comment on Some Recent Proposals of Carnap," in J. Hintikka (ed.), *Rudolf Carnap, Logical Empiricist: Materials and Perspectives*. Dordrecht, Netherlands: Reidel, 87–94.
Keynes, J. M. (1921), *A Treatise on Probability*. London: Macmillan & Co.
Lewis, C. I. (1918), *A Survey of Symbolic Logic*. Berkeley and Los Angeles: University of California Press.
Putnam, H. (1994), "Comments and Replies," in P. Clark and B. Hale (eds.), *Reading Putnam*. Oxford: Blackwell, 242–295.
Quine, W. V. O. (1951), "Two Dogmas of Empiricism," *Philosophical Review* 60: 20–43.
Richardson, A. (1998), *Carnap's Construction of the World*. Cambridge: Cambridge University Press.
Roozeboom, W. (1960), "A Note on Carnap's Meaning Criterion," *Philosophical Studies* 11: 33–38.
Sarkar, S. (1992), "'The Boundless Ocean of Unlimited Possibilities': Logic in Carnap's *Logical Syntax of Language*," *Synthese* 93: 191–237.
——— (2003), "Husserl's Role in Carnap's *Der Raum*," in T. Bonk (ed.), *Language, Truth and Knowledge: Contributions to the Philosophy of Rudolf Carnap*. Dordrecht, Netherlands: Kluwer, 179–190.
Sauer, W. (1985), "Carnap's '*Aufbau*' in Kantianishcer Sicht," *Grazer Philosophische Studien* 23: 19–35.
Savage, C. W. (2003), "Carnap's *Aufbau* Rehabilitated," in T. Bonk (ed.), *Language, Truth and Knowledge: Contributions to the Philosophy of Rudolf Carnap*. Dordrecht, Netherlands: Kluwer, 79–86.
Shimony, A. (1992), "On Carnap: Reflections of a Metaphysical Student," *Synthese* 93: 261–274.
Stegmüller, W. (1972), "Homage to Rudolf Carnap," in R. C. Buck and R. S. Cohen (eds.), *PSA 1970*. Dordrecht, Netherlands: Reidel, lii–lxvi.
Stein, H. (1992), "Was Carnap Entirely Wrong, After All?" *Synthese* 93: 275–295.

See also **Analyticity; Confirmation Theory; Cognitive Significance; Conventionalism; Hahn, Hans; Hempel, Carl Gustav; Induction, Problem of; Inductive Logic; Instrumentalism; Logical Empiricism; Neurath, Otto; Phenomenalism; Protocol sentences; Quine, Willard Van Orman; Reichenbach, Hans; Scientific Realism; Schlick, Moritz; Space-Time; Verifiability; Vienna Circle**

CAUSALITY

Arguably no concept is more fundamental to science than that of causality, for investigations into cases of existence, persistence, and change in the natural world are largely investigations into the causes of these phenomena. Yet the metaphysics and epistemology of causality remain unclear. For example, the ontological categories of the causal relata have been taken to be objects (Hume [1739] 1978), events (Davidson 1967), properties (Armstrong 1978), processes (Salmon 1984), variables (Hitchcock 1993), and facts (Mellor 1995). (For convenience, causes and effects will usually be understood as *events* in what follows.) Complicating matters, causal relations may be singular (''Socrates' drinking hemlock caused his death'') or general (''Drinking hemlock causes death''); hence the relata might be tokens (e.g., instances of properties) or types (e.g., types of events) of the category in question. Other questions up for grabs are: Are singular causes metaphysically and/or epistemologically prior to general causes or vice versa (or neither)? What grounds the intuitive asymmetry of the causal relation? Are macrocausal relations reducible to microcausal relations? And perhaps most importantly: Are causal facts (e.g., the holding of causal relations) reducible to noncausal facts (e.g., the holding of certain spatiotemporal relations)?

Some Issues in Philosophy of Causality: The Varieties of Causation

Causes can apparently contribute to effects in a variety of ways: by being background or standing conditions, ''triggering events,'' omissions, factors that enhance or inhibit effects, factors that remove a common preventative of an effect, etc. Traditionally accounts of causation have focussed on triggering events, but contemporary accounts are increasingly expected to handle a greater range of this diversity.

There may also be different notions of cause characteristic of the domains of different sciences (see Humphreys 1986; Suppes 1986): The seemingly indeterministic phenomena of quantum physics may require treatment different from either the seemingly deterministic processes of certain natural sciences or the ''quasi-deterministic'' processes characteristic of the social sciences, which are presumed to be objectively deterministic but subjectively uncertain. Another difference lies in the distinction between teleological (intentional, goal-oriented) and nonteleological causality: While the broadly physical sciences tend not to cite motives and purposes, the plant, animal, human, and social sciences often explicitly do. Contemporary treatments of teleological causality generally aim at avoiding the positing of anything like entelechies or ''vital forces'' (of the sort associated with nineteenth-century accounts of biology), and also at avoiding taking teleological goals to be causes that occur after their effects (see Salmon 1989, sec. 3.8, for a discussion). In Wright's (1976) account of ''consequence etiology,'' teleological behaviors (e.g., stalking a prey) are not caused by future catchings (which, after all, might not occur), but rather by the fact that the behavior in question has been often enough successful in the past that it has been evolutionarily selected for. Experience or inferences may also be operative in guiding behavior. While teleological causes raise interesting questions for the causal underpinnings of behavior (especially concerning whether a naturalistically acceptable account of intentionality can be given), the focus in what follows will be on nonteleological causality, reflecting the primary concern of contemporary philosophers of causation.

Singular vs. General Causation

Is all singular causation ultimately general? Different answers reflect different understandings of the notion of 'production' at issue in the platitude ''Causes produce their effects.'' In **generalist** (or covering-law) accounts (see the sections on ''Hume and Pearson: Correlation, Not Causation'' and ''Contemporary Philosophical Accounts of Causality: Generalist Accounts'' below), causal production is a matter of law: Roughly, event C causes event E just in case C and E are instances of terms in a law connecting events of C's type with events of E's type. The generalist interpretation is in part motivated by the need to ground inductive reasoning: Unless causal relations are subsumed by causal laws, one will be unjustified in inferring that events of C's type will, in the future, cause events of E's type. Another motivation stems from thinking that

identifying a sequence of events as causal requires identifying the sequence as falling under a (possibly unknown) law.

Alternatively, **singularists** (see ''Singularist Accounts'' below) interpret causal production as involving a singular causal process (variously construed) that is metaphysically prior to laws. Singularists also argue for the epistemological priority of singular causes, maintaining that one can identify a sequence as causal without assuming that the sequence falls under a law, even when the sequence violates modal presuppositions (as in Fair's [1979] case: Intuitively, one could recognize a glass's breaking as causal, even if one antecedently thought glasses of that type were unbreakable).

Counterfactual accounts (see ''Counterfactual Accounts'' below) analyze singular causes in terms of counterfactual conditionals (as a first pass, event C causes event E just in case if C had not occurred, then E would not have occurred). Whether a counterfactual account should be considered singularist, however, depends on whether the truth of the counterfactuals is grounded in laws connecting types of events or in, e.g., propensities (objective single-case chances) understood as irreducible to laws. Yet another option is to deny that either singular or general causes are reducible to the other, and go on to give independent treatments of each type (as in Sober 1984).

Reduction vs. Nonreduction

There are at least three questions of reducibility at issue in philosophical accounts of causation, which largely cut across the generalist/singularist distinction. The first concerns whether causal facts (e.g., the holding of causal relations) are reducible to noncausal facts (e.g., the holding of certain spatiotemporal relations). Hume's generalist reduction of causality (see ''Hume and Pearson: Correlation, Not Causation'' below) has a projectivist or antirealist flavor: According to Hume, the seeming ''necessary connexion'' between cause and effect is a projection of a psychological habit of association between ideas, which habit is formed by regular experience of events of the cause type being spatially contiguous and temporally prior to events of the effect type. Contemporary neo Humeans (see ''Hempel: Explanation, Not Causation'' and ''Probabilistic Relevance Accounts'' below) dispense with Hume's psychologism, focusing instead on the possibility of reducing causal relations and laws to objectively and noncausally characterized associations between events. (Whether such accounts are appropriately deemed antirealist is a matter of dispute, one philosopher's reductive elimination being another's reductive introduction.) By way of contrast, nonreductive generalists (often called **realists**—see ''Causal Powers, Capacities, Universals, Forces'' below) take the modally robust causal connection between event types to be an irreducible feature of reality (see Realism). Singularists also come in reductive or realist varieties (see ''Singularist Accounts'' below).

A second question of reducibility concerns whether a given account of causation aims to provide a conceptual analysis of the concept (hence to account for causation in bizarre worlds, containing magic, causal action at a distance, etc.) or aims to account for the causal relation in the actual world, in terms of physically or metaphysically more fundamental entities or processes. These different aims make a difference in what sort of cases and counterexamples philosophers of causation take to heart when developing or assessing theories. A common intermediate methodology focuses on central cases, leaving the verdict on far-fetched cases as ''spoils for the victor.''

A third question of reducibility concerns whether macrocausal relations (holding between entities, or expressed by laws, in the special sciences) are reducible to microcausal relations (holding between entities, or expressed by laws, in fundamental physics). This question arises from a general desire to understand the ontological and causal underpinnings of the structured hierarchy of the sciences, and from a need to address, as a special case, the ''problem of mental causation,'' of whether and how mental events (e.g., a feeling of pain) can be causally efficacious vis-à-vis certain effects (e.g., grimacing) that appear also to be caused by the brain events (and ultimately, fundamental physical events) upon which the mental events depend.

Causal reductionists (Davidson 1970; Kim 1984) suggest that mental events (more generally, macro-level events) are efficacious in virtue of supervening on (or being identical with) efficacious physical events. Many worry, however, that these approaches render macro-level events causally irrelevant (or ''epiphenomenal''). Nonreductive approaches to macro-level causation come in both physicalist and nonphysicalist varieties (Wilson 1999 provides an overview; see Physicalism). Some physicalists posit a relation (e.g., the determinable/determination relation or proper parthood) between macro- and micro-level events that entails that the set of causal powers of a given macrolevel event M (roughly, the set of causal interactions that the event, in appropriate circumstances, could enter into) is a proper subset of those of the micro-level event P upon which M

depends. In this ''proper subset'' strategy, the fact that the sets of causal powers are different provides some grounds for claiming that *M* is efficacious in its own right, but since each individual causal power of *M* is identical with a causal power of *P*, the two events are not in causal competition. In another nonreductive strategy—**emergentism**—the causal efficacy of at least some macro-level events (notably, mental events) is due to their having genuinely new causal powers not possessed by the physical events on which the mental events depend (see Emergence). When the effect in question is physical, such powers violate the causal closure of the physical (the claim that every physical effect has a fully sufficient physical cause); but such a violation arguably is not at odds with any cherished scientific principles, such as conversation laws (see McLaughlin 1992).

Features of Causality: Asymmetry, Temporal Direction, Transitivity

Intuitively, causality is asymmetric: If event *C* causes event *E*, then *E* does not cause *C*. Causality also generally proceeds from the past to the future. How to account for these data remains unclear. The problem of explaining asymmetry is particularly pressing for accounts that reductively analyze causality in terms of laws of association, for it is easy to construct cases in which the laws are reversible but the causation is not (see, for example, Bromberger's [1962] case in which the height *h* of a flagpole is correlated with the length *l* of the shadow it casts, and vice versa, and intuitively *h* causes *l* but *l* does not cause *h*). Both the asymmetry and the temporal direction of causality can be accommodated (as in Hume) by stipulatively identifying causal with temporal asymmetry: Causes differ from their effects in being prior to their effects. But this approach rules out simultaneous and backward causation, which are generally taken to be live possibilities; and it also rules out reducing the direction of time to the (general) direction of causation, which some (e.g., Reichenbach 1956) have wanted to do. The approach is also methodologically unsatisfactory, insofar as it fails to provide a basis for a unified resolution of problems facing associationist accounts (e.g., the problems of joint effects and of preemption discussed in ''Hume and Pearson: Correlation, Not Causation'' below).

Other strategies use physical processes to explain the temporal direction of causation. Reichenbach appealed to the direction of ''conjunctive forks'': processes in which a common cause produces joint effects and, in accordance with what Reichenbach called ''the principle of the common cause,'' the probabilistic dependence of the effects on each other is ''screened off''—goes away—when the common cause is taken into account. Such forks are, he claimed, always (or nearly always) open to the future and closed to the past. Others have suggested that the direction of causation is fixed by the direction of increasing entropy, or (more speculatively) by the direction of quantum collapse of the wave packet. Alternatively, Price (1992) suggests that human experience of manipulating causes provides a basis for the (projected) belief that causality is forward directed in time (see ''Counterfactuals and Manipulability'' below). All these accounts nonstipulatively explain the usual temporal direction of causal processes. But neither stipulative nor nonstipulative appeals to temporal direction seem to explain the asymmetry of causation, which intuitively has more to do with causes producing their effects (in some robust sense of 'production') than with causes being prior to their effects. Nonreductive accounts in which causality involves manifestations of powers or transfers of energy (or other conserved quantities) may be better situated to provide the required explanation, if such manifestations or transfers can be understood as directed (which remains controversial).

Another feature commonly associated with causality is transitivity: If *C* causes *D*, and *D* causes *E*, then *C* causes *E*. This assumption has come into question of late, largely due to the following sort of case (see Kvart 1991): A man's finger is severed in a factory accident; a surgeon reattaches the finger, which afterward becomes perfectly functional. The accident caused the surgery, and the surgery caused the finger's functionality; but it seems odd to say that the accident caused the finger's functionality. The precise bearing of such cases on the transitivity claim remains unclear (see Hall 2000 for a discussion).

Challenges to Causality: Galileo, Newton, and Maxwell—How, Not Why

From the ancient through modern periods, accounts of natural phenomena proceeded by citing the powers and capacities of agents, bodies, and mechanisms to bring about effects (see Hankinson 1998; Clatterbaugh 1999). Galileo's account of the physics of falling bodies initiated a different approach to scientific understanding, which became a matter of determining how certain measurable quantities were functionally correlated (the ''how'' of things, or the kinematics), as opposed to determining the causal mechanisms responsible for these

correlations (the ''why'' of things, or the dynamics). This descriptive approach enabled scientific theories to be formulated with relatively high precision, which in turn facilitated predictive and retrodictive success; by way of contrast, explanations in terms of (often unobservable) causal mechanisms came to be seen as explanatorily otiose at best and unscientific at worst.

Newton's famous claim in the *Principia* ([1687] 1999), ''Hypotheses non fingo'' (''I frame no hypotheses''), regarding gravitation's ''physical causes and seats'' is often taken as evidence that he advocated a descriptivist approach (though he speculated at length on the causes of gravitational forces in the *Optics*). And while Maxwell drew heavily upon Faraday's qualitative account of causally efficacious electromagnetic fields (and associated lines of force) in the course of developing his theories of electricity and magnetism, he later saw such appeals to underlying causes as heuristic aids that could be dropped from the final quantitative theory.

Such descriptivist tendencies have been encouraged by perennial worries about the metaphysical and epistemological presuppositions of explicitly causal explanations (see ''Causal Powers, Capacities, Universals, Forces'' below) and the concomitant seeming availability of eliminativist or reductivist treatments of causal notions in scientific laws. For example, Russell (1912) influentially argued that since the equations of physics do not contain any terms explicitly referring to causes or causal relations and moreover (in conflict with the presumed asymmetry of causality) appear to be functionally symmetric (one can write $a = f/m$ as well as $f = ma$), causality should be eliminated as ''a relic of a bygone age.'' Jammer (1957) endorsed a view in which forcebased dynamics is a sophisticated form of kinematics, with force terms being mere ''methodological intermediaries'' enabling the convenient calculation of quantities (e.g., accelerations) entering into descriptions. And more recently, van Fraassen (1980) has suggested that while explanations going beyond descriptions may serve various pragmatic purposes, these have no ontological or causal weight beyond their ability to ''save the phenomena.''

Whether science really does, or should, focus on the (noncausally) descriptive is, however, deeply controversial. Galileo himself sought for explanatory principles going beyond description (see Jammer 1957 for a discussion), and notwithstanding Newton's professed neutrality about their physical seats, he took forces to be the ''causal principle[s] of motion and rest.'' More generally, notwithstanding the availability of interpretations of scientific theories as purely descriptive, there are compelling reasons (say, the need to avoid a suspect action at a distance) for taking the causally explanatory posits of scientific theories (e.g., fields and forces) ontologically seriously. The deeper questions here, of course, concern how to assess the ontological and causal commitments of scientific theories; and at present there is no philosophical consensus on these important matters. In any case it is not enough, in assessing whether causes are implicated by physical theories, to note that terms like ''cause'' don't explicitly appear in the equations of the theory, insofar as the commitments of a given theory may transcend the referents of the terms appearing in the theory, and given that many terms—force, charge, valence—that do appear are most naturally defined in causal terms ('force,' for example, is usually defined as that which causes acceleration). It is also worth noting that the apparent symmetry of many equations, as well as the fact that cause terms do not explicitly appear in scientific equations, may be artifacts of scientists' using the identity symbol as an all-purpose connective between functional quantities, which enables the quantities to be manipulated using mathematical techniques but is nonetheless implicitly understood as causally directed, as in $f = ma$.

Nor does scientific practice offer decisive illumination of whether scientific theorizing is or is not committed to causal notions: As with Maxwell, it remains common for scientists to draw upon apparently robustly causal notions when formulating or explaining a theory, even while maintaining that the theory expresses nothing beyond descriptive functional correlations of measurable quantities. Perhaps it is better to attend to what scientists do rather than what they say. That they rarely leave matters at the level of descriptive laws linking observables is some indication that they are not concerned with just the ''how'' question—though, to be sure, the tension between descriptive and causal/explanatory questions may recur at levels below the surface of observation.

Hume and Pearson: Correlation, Not Causation

The Galilean view of scientific understanding as involving correlations among measurable quantities was philosophically mirrored in the empiricist view that all knowledge (and meaning) is ultimately grounded in sensory experience (see Empiricism). The greatest philosophical challenge to causality came from Hume, who argued that there is no

experience of causes being efficacious, productive, or powerful vis-à-vis their effects; hence "we only learn by experience the frequent *conjunction* of objects, without being ever able to comprehend any thing like *connexion* between them" ([1748] 1993, 46). In place of realistically interpreted "producing" theories of causation (see Strawson 1987 for a taxonomy), Hume offered the first regularity theory of causation, according to which event C causes event E just in case C and E occur, and events of C's type have (in one's experience) been universally followed by, and spatially contiguous to, events of E's type. (As discussed, such constant conjunctions were the source of the psychological imprinting that was, for Hume, the ultimate locus of causal connection.) Hume's requirement of contiguity may be straightforwardly extended to allow for causes to produce distant effects, via chains of spatially contiguous causes and effects.

Hume's requirement of universal association is not sufficient for causation (night always follows day but night does not cause day). Nor is Hume's requirement necessary, for even putting aside the requirement that one experience the association in question, there are many causal events that happen only once (e.g., the big bang). The immediate move of neo-Humeans (e.g., Hempel and Oppenheim 1948; Mackie 1965) is to understand the generalist component of the account in terms of laws of nature, which express the lawful sufficiency of the cause for the effect, and where (reflecting Hume's reductive approach) the sufficiency is not to be understood as grounded in robust causal production. One wonders, though, what is grounding the laws in question if associations are neither necessary nor sufficient for their holding and robust production is not allowed to play a role. If the laws are grounded in brute fact, it is not clear that the reductive aim has been served (but see the discussion of Lewis's account of laws, below).

In any case, neo-Humean accounts face several problems concerning events that are inappropriately deemed causes ("spurious causes"). One is the problem of joint effects, as when a virus causes first a fever, and then independently causes a rash: Here the fever is lawfully sufficient for, hence incorrectly deemed a cause of, the rash. Another involves violations of causal asymmetry: Where events of the cause's type are lawfully necessary for events of the effect's type, the effect will be lawfully sufficient for (hence inappropriately deemed a cause of) the cause. Cases of preemption also give rise to spurious causes: Suzy's and Billy's rockthrowings are each lawfully sufficient for breaking the bottle; but given that Suzy's rock broke the bottle (thereby preempting Billy's rock from doing so), how is one to rule out Billy's rockthrowing as a cause?

The above cases indicate that lawful sufficiency alone does not satisfy causality. One response is to adopt an account of events in which these are finely individuated, so that, for example, the bottlebreaking resulting from Suzy's rockthrowing turns out to be of a different event type than a bottlebreaking resulting from Billy's rockthrowing (in which case Billy's rockthrowing does not instantiate a rockthrowing–bottlebreaking law, and so does not count as a cause). Another response incorporates a proviso that the lawful sufficiency at issue is that of the circumstances (as in Mackie's "INUS" condition account, in which a cause is an *i*nsufficient but *n*ecessary part of a condition that is, in the circumstances, *u*nnecessary but *s*ufficient for the effect).

Nor is lawful sufficiency alone necessary for causality, as the live possibility of irreducibly probabilistic causality indicates. This worry is usually sidestepped by a reconception of laws according to which these need express only some lawlike pattern of dependence; but this reconception makes it yet more difficult for regularity theorists to distinguish spurious from genuine causes and laws (see "Probabilistic Relevance Accounts" below for developments). One neo-Humean response is to allow certain *a priori* constraints to enter into determining what laws there are in a world (as in the "best system" theory of laws of Lewis 1994) in which the laws are those that systematize the phenomena with the best combination of (predictive) strength and (formal) simplicity, so as to accommodate probabilistic (and even uninstantiated) laws (see also "Hempel: Explanation, Not Causation" below).

The view that causation is nothing above (appropriately complex) correlations was widespread following the emergence of social statistics in the nineteenth century and was advanced by Karl Pearson, one of the founders of modern statistics, in 1890 in *The Grammar of Science*. Pearson's endorsement of this view was, like Hume's, inspired by a rejection of causality as involving mysterious productive powers, and contributed to causes (as opposed to associations) being to a large extent expunged from statistics and from the many sciences relying upon statistics. An intermediate position between these extremes, according to which causes are understood to go beyond correlations but are not given any particular metaphysical interpretation (in particular, as involving productive powers), was advanced by the evolutionary biologist Sewall Wright, the inventor of path

AU:4

analysis. Wright (1921) claimed that path analysis not only enabled previously known causal relations to be weighted, but moreover enabled the testing of causal hypotheses in cases where the causal relations were (as yet) unknown. Developed descendants and variants of Wright's approach have found increasing favor of late (see "Bayesian Networks and Causal Models" below), contributing to some rehabilitation of the notion of causality in the social sciences.

Hempel: Explanation, Not Causation

Logical empiricists were suspicious of causation understood as a metaphysical connection in nature; instead they located causality in language, interpreting causal talk as talk of explanation (see Salmon 1989 for a discussion). On Hempel and Oppenheim's (1948) influential D-N (deductive-nomological) model of scientific explanation, event C explains event E just in case a statement expressing the occurrence of E is the conclusion of an argument with premises, one of which expresses the holding of a universal generalization to the effect that events of C's type are associated with events of E's type and another of which expresses the fact that C occurred (see Explanation). Imposing certain requirements on universal generalizations (e.g., projectibility) enabled the D-N account to avoid cases of spurious causation due to accidental regularities (S's being a screw in Smith's car does not explain why S is rusty, even if all the screws in Smith's car are rusty). Hempel suggested various strategies for how the D-N account might avoid admitting spurious causes (explanations) associated with cases of asymmetry and preemption. For example, if one is willing to let causal asymmetry depend on temporal asymmetry, one can avoid admitting the length of the shadow as explaining the height of the flagpole by noting that the length of the shadow at T depends on the height of the flagpole at $T-\varepsilon$, while no such relationship holds between the height of the flagpole at T and the length of the shadow at $T-\varepsilon$.

AU:5

AU:6

To accommodate the possibility of irreducibly probabilistic associations, as well as explanations (characteristic of the social sciences) proceeding under conditions of partial uncertainty, Hempel (1965) proposed an inductive-statistical (I-S) model, in which event C explains event E if C occurs and it is an inductively grounded law that the probability of an event of type E given an event of type C is high (see Explanation; Inductive Logic). This account is subject to counterexamples in which a cause produces an effect but with a low probability, as in Scriven's case (discussed by him prior to Hempel's extension, and developed in Scriven 1975), where the probability of paresis given syphilis is low, but when paresis occurs, syphilis is the reason. A similar point applies to many quantum processes. Such cases gave rise to two different approaches to handling probabilistic explanation (or causation, by those inclined to accept this notion). One approach (see Railton 1978) locates probabilistic causality in propensities; the other (see "Probabilistic Relevance Accounts" below) in more sophisticated probabilistic relations.

At this point the line between accounts that are reductive (in the sense of reducing causal to noncausal goings-on) and nonreductive, as well as the line between singularist and generalist accounts, begins to blur. For while a propensity-based account of causality initially looks nonreductive and singularist, some think that propensities can be accommodated in a sophisticated associationist account of laws; and while an account based on relations of probabilistic relevance initially looks reductive and generalist, whether it is so depends on how the probabilities are interpreted (as given by frequencies, irreducible propensities, etc.).

Contemporary Philosophical Accounts of Causality: Generalist Accounts

Probabilistic Relevance Accounts

A natural response to Scriven-type cases is to understand positive causal relevance in terms of probability raising (Suppes 1970): Event C causes event E just in case the probability of events of E's type is higher given events of C's type than without. (Other relevance relations, such as being a negative causal factor, can be defined accordingly.) A common objection to such accounts (see Rosen 1978) proceeds by constructing cases "the hard way," in which it seems that causes lower the probability of their effects (e.g., where a mishit golf ball ricochets off a tree, resulting in a hole-in-one; or where a box contains a radioactive substance S that produces decay particles but the presence of S excludes the more effective radioactive substance S′). Such cases can often be handled, however, by locating a neutral context (where the golf ball is not hit at all, or where no radioactive substance is in the box) relative to which events of the given type do raise the probability of events of the effect type.

AU:7

A more serious problem for probability-raising accounts is indicated by Simpson's paradox, according to which any statistical relationship between two variables may be reversed by including additional factors in the analysis. The "paradox"

reflects the possibility that a variable C can be positively correlated with a variable E in a population and yet C be negatively correlated with E in every partition of the population induced by a third variable X. Just this occurred in the Berkeley sex discrimination case. Relative to the population of all men and women applying to graduate school at the University of California, Berkeley, being male (C) was positively correlated with being admitted (E). But relative to every partition of the population containing the men and women applying to a particular department (X), this correlation was reversed. In this case the difference between the general and specific population statistics reflected the fact that while in every department it was easier for women to be admitted than men, women were more likely to apply to departments that were harder (for *everyone*) to get into. The general population statistic was thus confounded: Being a male, *simpliciter*, was not in fact causally relevant to getting into graduate school at UC Berkeley; rather (assuming no other confounding was at issue), applying to certain departments rather than others was what was relevant.

To use probabilistic accounts as a basis for testing hypotheses and making predictions—and especially in order to identify effective strategies (courses of action) in the social sciences and, indeed, in everyday life—statistical confounding needs to be avoided. Cartwright (1979) suggested that avoiding confounding requires that the relevant probabilities be assessed relative to background contexts within which all other causal factors (besides the variable C, whose causal relevance is at issue) are held fixed. (Of course, reductionists need to be able to specify these background contexts without appealing to distinctly causal factors; see below.) Opinions differ regarding whether events of type C must raise the probability of events of type E in at least one such context, in a majority of contexts, or in every such context. So one might take C to be a positive causal factor for E just in case $p(E \mid C \wedge X_i) \geq p(E \mid \overline{C} \wedge X_i)$ for all background contexts X_i, with strict inequality for at least one X_i. In this approach, smoking would be a positive causal factor for having lung cancer just in case smoking increases the chance of lung cancer in at least one background context and does not lower it in any background context.

Practically, Cartwright's suggestion has the disadvantage that one is frequently not in a position to control for all alternative causal factors (though in some circumstances one can avoid having to do this; see "Bayesian Networks and Causal Models" below). Philosophically, the requirement threatens reductive versions of probabilistic accounts with circularity. Attempts have been made to provide a noncircular means of specifying the relevant background contexts (e.g., Salmon 1984), but it is questionable whether these attempts succeed, and many are presently prepared to agree with Cartwright: "No causes in, no causes out."

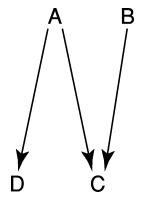
As mentioned, probabilistic accounts may or may not be reductive, depending on whether the probabilities at issue are understood as grounded in associations (as in Suppes 1970) or else in powers, capacities, or propensities (as in Humphreys 1989 and Cartwright 1989). In the latter interpretation, further divisions are introduced: If the propensities are taken to be irreducible to laws (as in Cartwright's account), then the associated probabilistic relevance account is more appropriately deemed singularist. Complicating the taxonomy here is the fact that most proponents of probabilistic accounts are not explicit as regards what analysis should be given of the probabilities at issue.

Bayesian Networks and Causal Models

Philosophical worries concerning whether statistical information adequately tracks causal influence are echoed in current debates over the interpretation of the statistical techniques used in the social sciences. As noted above ("Hume and Pearson: Correlation, Not Causation"), these techniques have frequently been interpreted as relating exclusively to correlations, but increasingly researchers in computer science, artificial intelligence, and statistics (see Spirtes, Glymour, and Scheines 1993; Pearl 2000) have developed nonreductive interpretations of these approaches as encoding explicitly causal information and that appear to lead to improved hypothesis testing and prediction of effects under observation and intervention.

Another advantage claimed for such accounts is that they provide a means of avoiding confounding without imposing the often impracticable requirement that the relevant probabilities be assessed against background contexts taking into account all causal factors, it rather being sufficient to take into account all *common* causal factors. As a simple illustration, suppose that A and B are known to causally influence C, as in Figure 1.

To judge whether D influences C, Cartwright generally recommends holding fixed both A and B, while Spirtes et al. (1993) instead recommend holding fixed only A. Cartwright allows, however, that attention to just common causal factors is possible when the causal Markov condition (discussed below) holds. As will be seen shortly, this



AU:11 Fig. 1.

condition must hold in order to implement the causal modeling approach. So, the restriction to common causal factors is not really an advantage over Cartwright's account.

In the causal modeling approach, one starts with a set of variables (representing properties) and a probability distribution over the variables. This probability distribution, partially interpreted with prior causal knowledge, is assumed to reflect a causal structure (a set of laws expressing the causal relations between the variables, which laws may be expressed either graphically or as a set of structurally related functional equations). Given that certain conditions (discussed below) hold between the probabilities and the causal structure, algorithmic techniques are used to generate the set of all causal structures consistent with the probabilities and the prior causal knowledge. Techniques also exist for extracting information regarding the results of interventions (corresponding to manipulations of variables).

AU:13

While causal modeling approaches may lead to improved causal inference concerning complex systems (in which case they are of some epistemological interest), it is unclear what bearing they have on the metaphysics of causality. Spirtes et al. (1993) present their account not so much as an analysis of causality as a guide to causal inference. Pearl (2000), however, takes the appeal to prior causal intuitions to indicate that causal modeling approaches are nonreductive (and moreover based on facts about humans' cognitive capacities to make effective causal inferences in simple cases). In any case, the potential of causal models to provide a basis for a general theory of causality is limited by the fact that certain strong conditions need to be in place in order for the algorithms to be correctly applied.

One of these is the **causal Markov condition** (of which Reichenbach's [1956] "principle of the common cause" was a special case), which says that once one conditions on the complete set P_v of "causal parents" (direct causes) of a variable V, V will be probabilistically independent of all other variables except V's descendants; that is, for all variables X, where X is not one of V's descendants, $P(V \mid P_v \wedge X) = P(V \mid P_v)$. (In particular, where V is a joint effect, conditioning on the causal parents of V screens off the probabilistic influence of the other joint effects on V.) Here again there is the practical problem that in the social sciences, where the approaches are supposed to be applicable, one is often not in a position to specify states with sufficient precision to guarantee that the condition is met. A metaphysical problem is that (contrary to Reichenbach's apparent assumption that the condition holds in all cases involving a common cause of joint effects) the causal Markov condition need not hold in cases of probabilistic causation: When a particle may probabilistically decay either by emitting a high-energy electron and falling into a lowenergy state or by emitting a low-energy electron and falling into a different energy state, the joint effects in either case will not be probabilistically independent of each other, even conditioning on the cause; and certain cases of macrocausation appear also to violate the condition.

AU:14 AU:15

A second assumption of the causal modeling technique is what Spirtes et al. (1993) call "faithfulness" (also known as "stability" in Pearl 2000), according to which probabilistic dependencies faithfully reveal causal connections. In particular, if Y is probabilistically independent of X, given X's parents, then X is assumed not to cause Y. Again, this condition cannot be assumed to hold in all cases, since some variables (properties) may sometimes prevent and sometimes produce an effect (as when birth control pills are a cause of thrombosis yet also prevent thrombosis, insofar as pregnancy causes thrombosis and the pills prevent pregnancy). In circumstances where the positive and negative contributions of X to Y are equally effective, the probabilistic dependence of effect on cause may cancel out, and thus X may inappropriately be taken not to be causally relevant to Y.

Causal Powers, Capacities, Universals, Forces

As mentioned, some proponents of probabilistic relevance accounts endorse metaphysical interpretations of the probabilities at issue. Such positions fall under the broader category of nonreductive ("realist") **covering-law** theories, in which laws express (or are grounded in) more than mere associations. The job such accounts face is to provide an alternative basis for causal laws. Among other possibilities, these bases are taken to be relations of necessitation or "probabilification" among universals (Dretske 1977; Tooley 1977; Armstrong 1978),

(law-based) capacities or powers associated with objects or properties (Shoemaker 1980; Martin 1993), or fundamental forces or interactions (Bohm 1957; Strawson 1987).

Such accounts sidestep many of the problems associated with reductive coveringlaw accounts. Since laws are not just a matter of association, a realist has the means to deny, in the virus–fever–rash case, that there is a law connecting fevers with rashes; similarly, in cases of preemption a realist may claim that, for example, Billy's rockthrowing and the bottle's breaking did not instance the law in question (even without endorsing a fine-grained account of event individuation). Of course, much depends here on the details of the proposed account of laws. In the Dretske-Tooley-Armstrong account, causal laws are contingent, brute relations between universals. Some find this less than satisfying from a realist point of view, insofar as it is compatible with, for example, the property of having spin 1 bestowing completely different causal powers (say, all those actually bestowed by having spin $\frac{1}{2}$) on its possessing particulars. Other realists are more inclined to see the nature of properties and particulars as essentially dependent on the causal laws that actually govern them, a view that is quite plausible for scientific entities: "[C]ausal laws are not like externally imposed legal restrictions that, so to speak, merely limit the course of events to certain prescribed paths . . . . [T]he causal laws satisfied by a thing . . . are inextricably bound up with the basic properties of the thing which helps to define what it is" (Bohm 1957, 14).

The primary problem facing realist accounts is that they require accepting entities and relations (universals, causal powers, forces) that many philosophers and scientists find metaphysically obscure and/or epistemologically inaccessible. How one evaluates these assessments often depends on one's other commitments. For example, many traditional arguments against realist accounts (e.g., Hume's arguments) are aimed at showing that these do not satisfy a strict epistemological standard, according to which the warranted posit of a contingent entity requires that the entity be directly accessible to experience (or a construction from entities that are so accessible). But if inference to the existence of an unexperienced entity (as the best explanation of some phenomena) is at least sometimes an acceptable mode of inference, such a strict epistemological standard (and associated arguments) will be rejected; and indeed, positive arguments for contemporary realist accounts of causality generally proceed via such inferences to the best explanation—often of the patterns of association appealed to by reductivist accounts.

AU:16

Singularist Accounts

Singularists reject the claim that causes follow laws in the order of explanation, but beyond this there is considerable variety in their accounts. *Contra* Hume, Anscombe (1971) takes causation to be a (primitive) relation that may be observed in cuttings, pushings, fallings, etc. It is worth noting that a primitivist approach to causality is compatible with even a strict empiricism (compare Hume's primitivist account of the resemblance relation). The empiricist Ducasse (1926) also locates causation in singular observation, but nonprimitively: A cause is the change event observed to be immediately prior and spatiotemporally contiguous to an effect event. While interesting in allowing for a nonassociative, nonprimitivist, empiricist causality, Ducasse's account is unsatisfactory in allowing only the coarse-grained identification of causes (as some backward temporal segment of the entire observed change); hence it fails to account for most ordinary causal judgments. Note that singularists basing causation on observation need not assert that one's knowledge of causality proceeds only via observations of the preferred sort; they rather generally maintain that such experiences are sufficient to account for one's acquiring the concept of causation, then allow that causation need not be observed and that confirming singular causal claims may require attention to associations.

Another singularist approach takes causation to be theoretically inferred, as that relation satisfying (something like) the Ramsey sentence consisting of the platitudes about causality involving asymmetry, transitivity, and so on (see Tooley 1987). One problem here is that, as may be clear by now, such platitudes do not seem to uniformly apply to all cases. Relatedly, one may wonder whether they are consistent; given the competing causal intuitions driving various accounts of causality, it would be surprising if they were.

Finally, a wide variety of singularist accounts understand causality in terms of singular processes. Such accounts are strongly motivated by the intuition that in a case of preemption such as that of Suzy and Billy, what distinguishes Suzy's throw as a cause is that it initiates a process ending in the bottle breaking, while the process initiated by Billy's throw never reaches completion (see Menzies 1996 for a discussion). Commonly, process singularists attempt (like Ducasse) to provide a nonprimitivist causality that is both broadly empiricist, in not

appealing to any properly metaphysical elements, and nonassociationist, in recognition of the diffculties that associationist accounts have (both with preemption and with distinguishing genuine causality from accidental regularity). Hence they typically fill in the "process" intuition by identifying causality with fundamental physical processes, including transfers or interactions, as in Fair's (1979) account of causation as identical with the transfer of energy momentum, Salmon's (1984) "mark transmission" account, and Dowe's (1992) account in which the transfer of any conserved quantity will suffice.

An objection to the claim that physical processes are sufficient for causality is illustrated by Cartwright's (1979) case of a plant sprayed with herbicide that improbably survives and goes on to flourish (compare also Kvart's 1991 finger-severing case, discussed previously). While transfers and interactions of the requisite sort can be traced from spraying to flourishing, intuitively the former did not cause the latter; however, accepting that the spraying did cause the flourishing may not be an overly high price to pay. A deeper worry concerns the epistemological question of how accounts of physical process link causation understood as involving theoretical relations or processes of fundamental physics with causation as ordinarily experienced. Fair suggests that ordinary experience involves macroprocesses, which are in turn reducible to the relevant physical processes; but even supposing that such reductions are in place, ordinary causal judgments do not seem to presuppose them.

Counterfactual Accounts

Counterfactual accounts of causality take as their starting point the intuition that a singular cause makes an important difference in what happens. As a first pass, C causes E (where C and E are actually occurring events) only if, were C not to occur, then E would not occur. As a second pass, C causes E only if C and E are connected by a chain of such dependencies (see Lewis 1973), so as to ensure that causation is transitive (that is, causation is the "ancestral," or transitive closure, of counterfactual dependence). In addition to the requirement of counterfactual necessity of causes for effects, counterfactual accounts also commonly impose a requirement of counterfactual sufficiency of causes for effects: If C were to occur, then E would occur. Insofar as counterfactual accounts are standardly aimed at reducing causal to noncausal relations, and given plausible assumptions concerning evaluation of counterfactuals, the latter requirement is satisfied just by C and E's actually occurring (which occurrences, as above, are assumed); hence standard counterfactual accounts do not have a nontrivial notion of counterfactual sufficiency. A nontrivial notion of counterfactual sufficiency can be obtained by appeal to nested counterfactuals (see Vihvelin 1995): C causes E only if, if neither C nor E had occurred, then (if C had occurred, then E would have occurred). AU:17

Problems, Events, and Backtrackers

While counterfactual accounts are often motivated by a desire to give a reductive account of causality that avoids problems with reductive covering-law accounts (especially those of joint effects and of preemption), it is unclear whether counterfactual accounts do any better by these problems. First, consider the problem of joint effects. Suppose a virus causes first a fever, then a rash, and that the fever and rash could only have been caused by the virus. It seems correct to reason in the following "backtracking" fashion: If the fever had not occurred, then the viral infection would not have occurred, in which case the rash would not have occurred. But then the counterfactual "If the fever had not occurred, the rash would not have occurred" turns out true, which here means that the fever causes the rash, which is incorrect. Proponents of counterfactual accounts have responses to these objections, which require accepting controversial accounts of the truth conditions for counterfactuals (see Lewis 1979). Even so, the responses appear not to succeed (see Bennett 1984 for a discussion).

Second, consider the problem of preemption. In the Suzy-Billy case, it seems correct to reason that if Suzy had not thrown her rock, then Billy's rock would have gotten through and broken the bottle. Hence the counterfactual "If Suzy's throw had not occurred, the bottlebreaking would not have occurred" turns out false; so Suzy's rockthrowing turns out not to be a cause, which is incorrect. In cases (as here) of so-called "early preemption," where it makes sense to suppose that there was an intermediate event D between the effect and the cause on which the effect depended, this result can be avoided: Although the breaking does not counterfactually depend on Suzy's rockthrowing, there is a chain of counterfactual dependence linking the breaking to Suzy's rockthrowing (and no such chain linking the breaking to Billy's), and so her throw does end up being a cause (and Billy's does not). But the appeal to an intermediate event seems ad hoc and in any case cannot resolve cases

of "late preemption." Lewis (developing an idea broached in Paul 1998) eventually responded to such cases by allowing that an event may be counted as a cause if it counterfactually influences the mode of occurrence of the effect (e.g., how or when it occurs), as well as if it counterfactually influences the occurrence of the effect, *simpliciter*.

Counterfactuals and Manipulability

Where counterfactual accounts may be most useful is in providing a basis for understanding or formalizing the role that manipulability plays in the concept of causation. One such approach sees counterfactuals as providing the basis for an epistemological, rather than a metaphysical, account of causation (see Pearl 2000 for a discussion). The idea here is that counterfactuals nicely model the role manipulability (actual or imagined) plays in causal inference, for a natural way to determine whether a counterfactual is true is to manipulate conditions so as to actualize the antecedent. Another approach takes counterfactuals to provide a basis for a generalist account of causal explanation (see Woodward 1997), according to which such explanations track stable or invariant connections and the notion of invariance is understood nonepistemologically in terms of a connection's continuing to hold through certain counterfactual (not necessarily human) "interventions." Whether the notion of manipulability is itself a causal notion, and so bars the reduction of causal to noncausal facts, is still an open question.

JESSICA WILSON

AU:18 **References**

Anscombe, G. E. M. (1971), *Causality and Determination*. Cambridge: Cambridge University Press.

Armstrong, David M. (1978), *Universals and Scientific Realism*, vol. 2: *A Theory of Universals*. Cambridge: Cambridge University Press.

Bennett, Jonathan (1984), "Counterfactuals and Temporal Direction," *Philosophical Review* 93: 57–91.

Bohm, David (1957), *Causality and Chance in Modern Physics*. London: Kegan Paul.

Bromberger, Sylvain (1962), "What Are Effects?" in Ronald J. Butler (ed.), *Analytical Philosophy*, vol. 1. Oxford: Oxford University Press.

Cartwright, Nancy (1979), "Causal Laws and Effective Strategies," *Noûs* 13: 419–438.

——— (1989), *Nature's Capacities and Their Measurement*. Oxford: Clarendon Press.

Clatterbaugh, Kenneth (1999), *The Causation Debate in Modern Philosophy, 1637–1739*. New York: Routledge.

Davidson, Donald (1967), "Causal Relations," *Journal of Philosophy* 64: 691–703. Reprinted in Davidson (2001), *Essays on Actions and Events*. Oxford: Oxford University Press, 149–162.

——— (1970), "Mental Events," in L. Foster and J. Swanson (eds.), *Experience and Theory*. Amherst: Massachusetts University Press. Reprinted in Davidson (2001), *Essays on Actions and Events*. Oxford: Oxford University Press, 207–224.

Dowe, Phil (1992), "Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory," *Philosophy of Science* 59: 195–216.

Dretske, Fred (1977), "Laws of Nature," *Philosophy of Science* 44: 248–268.

Ducasse, C. J. (1926), "On the Nature and Observability of the Causal Relation," *Journal of Philosophy* 23: 57–68.

Fair, David (1979), "Causation and the Flow of Energy," *Erkenntnis* 14: 219–250.

Hall, Ned (2000), "Causation and the Price of Transitivity," *Journal of Philosophy* 97: 198–222.

Hankinson, R. J. (1998), *Cause and Explanation in Ancient Greek Thought*. Oxford: Oxford University Press.

Hempel, Carl (1965), *In Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.

Hempel, Carl, and Paul Oppenheim (1948), "Studies in the Logic of Explanation," *Philosophy of Science* 15: 135–175.

Hitchcock, Christopher (1993), "A Generalized Probabilistic Theory of Causal Relevance," *Synthese* 97: 335–364.

Hume, David ([1739] 1978), *A Treatise of Human Nature*. Edited by L. A. Selby-Bigge. Oxford: Oxford University Press.

——— ([1748] 1993), *An Enquiry Concerning Human Understanding*. Edited by Eric Steinberg. Indianapolis: Hackett Publishing.

Humphreys, Paul (1986), "Causation in the Social Sciences: An Overview," *Synthese* 68: 1–12.

——— (1989), *The Chances of Explanation*. Princeton, NJ: Princeton University Press.

Jammer, Max (1957), *Concepts of Force*. Cambridge, MA: Harvard University Press.

Kim, Jaegwon (1984), "Epiphenomenal and Supervenient Causation," *Midwest Studies in Philosophy IX: Causation and Causal Theories*, 257–270.

——— (1993), *Supervenience and Mind: Selected Philosophical Essays*. Cambridge: Cambridge University Press.

Kvart, Igal (1991), "Transitivity and Preemption of Causal Relevance," *Philosophical Studies* LXIV: 125–160.

Lewis, David (1973), "Causation," *Journal of Philosophy* 70: 556–567.

——— (1979), "Counterfactual Dependence and Time's Arrow," *Noûs* 13:455–476.

——— (1983), *Philosophical Papers*, vol. 1. Oxford: Oxford University Press.

——— (1994), "Humean Supervenience Debugged," *Mind* 1: 473–490.

Mackie, John L. (1965), "Causes and Conditions," *American Philosophical Quarterly* 2: 245–264.

Martin, C. B. (1993), "Power for Realists," in Keith Cambell, John Bacon, and Lloyd Reinhardt (eds.), *Ontology, Causality, and Mind: Essays on the Philosophy of D. M. Armstrong*. Cambridge: Cambridge University Press.

McLaughlin, Brian (1992), "The Rise and Fall of British Emergentism," in Ansgar Beckerman, Hans Flohr, and Jaegwon Kim (eds.), *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*. Berlin: De Gruyter, 49–93.

Mellor, D. H. (1995), *The Facts of Causation*. Oxford: Oxford University Press.

Menzies, Peter (1996), "Probabilistic Causation and the Pre-Emption Problem," *Mind* 105: 85–117.
Newton, Isaac ([1687] 1999), *The Principia: Mathematical Principles of Natural Philosophy*. Translated by I. Bernard Cohen and Anne Whitman. Berkeley and Los Angeles: University of California Press.
Paul, Laurie (1998), "Keeping Track of the Time: Emending the Counterfactual Analysis of Causation," *Analysis* LVIII: 191–198.
Pearl, Judea (2000), *Causality*. Cambridge: Cambridge University Press.
Price, Huw (1992), "Agency and Causal Asymmetry," *Mind* 101: 501–520.
Railton, Peter (1978), "A Deductive-Nomological Model of Probabilistic Explanation," *Philosophy of Science* 45: 206–226.
Reichenbach, Hans (1956), *The Direction of Time*. Berkeley and Los Angeles: University of California Press.
Rosen, Deborah (1978), "In Defense of a Probabilistic Theory of Causality," *Philosophy of Science* 45: 604–613.
Russell, Bertrand (1912), "On the Notion of Cause," *Proceedings of the Aristotelian Society* 13: 1–26.
Salmon, Wesley (1984), *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.
——— (1989), "Four Decades of Scientific Explanation," in Philip Kitcher and Wesley Salmon (eds.), *Scientific Explanation*, 3–219.
Scriven, Michael (1975), "Causation as Explanation," *Noûs* 9: 238–264.
Shoemaker, Sydney (1980), "Causality and Properties," in Peter van Inwagen (ed.), *Time and Cause*. Dordrecht, Netherlands: D. Reidel, 109–135.
Sober, Elliott (1984), "Two Concepts of Cause," in Peter D. Asquith and Philip Kitcher (eds.), *PSA 1984*. East Lansing, MI: Philosophy of Science Association, 405–424.
Sosa, Ernest, and Michael Tooley (eds.) (1993), *Causation*. Oxford: Oxford University Press..
Spirtes, P., C. Glymour, and R. Scheines (1993), *Causation, Prediction, and Search*. New York: Springer-Verlag.
Strawson, Galen (1987), "Realism and Causation," *Philosophical Quarterly* 37: 253–277.
Suppes, Patrick (1970), *A Probabilistic Theory of Causality*. Amsterdam: North-Holland.
——— (1986), "Non-Markovian Causality in the Social Sciences with Some Theorems on Transitivity," *Synthese* 68: 129–140.
Tooley, Michael (1977), "The Nature of Laws," *Canadian Journal of Philosophy* 7: 667–698.
——— (1987), *Causation: A Realist Approach*. Oxford: Oxford University Press.
van Fraassen, Bas (1980), *The Scientific Image*. Oxford: Oxford University Press.
Vihvelin, Kadhri (1995), "Causes, Effects, and Counterfactual Dependence," *Australasian Journal of Philosophy* 73: 560–583.
Wilson, Jessica (1999), "How Superduper Does a Physicalist Supervenience Need to Be?" *Philosophical Quarterly* 49: 33–52.
Woodward, James (1997), "Explanation, Invariance, and Intervention," *Philosophy of Science* 64 (Proceedings): S26–S41.
Wright, Larry (1976), *Teleological Explanation*. Berkeley and Los Angeles: University of California Press.
Wright, Sewall (1921), "Correlation and Causation," *Journal of Agricultural Research* 20: 557–585.

AU:19

CHAOS THEORY

See **Prediction**

CHEMISTRY, PHILOSOPHY OF

Although many influential late-nineteenth- and early-twentieth-century philosophers of science were educated wholly or in part as chemists (Gaston Bachelard, Pierre Duhem, Emile Myerson, Wilhelm Ostwald, Michael Polanyi), they seldom reflected directly on the epistemological,

methodological, or metaphysical commitments of their science. Subsequent philosophers of science followed suit, directing little attention to chemistry in comparison with physics and biology despite the industrial, economic, and academic success of the chemical sciences. Scattered examples of philosophical reflection on chemistry by chemists do exist; however, philosophically sensitive historical analysis and sustained conceptual analysis are relatively recent phenomena. Taken together, these two developments demonstrate that chemistry addresses general issues in the philosophy of science and, in addition, raises important questions in the interpretation of chemical theories, concepts, and experiments.

Historians of chemistry have also raised a number of general philosophical questions about chemistry. These include issues of explanation, ontology, reduction, and the relative roles of theories, experiments, and instruments in the advancement of the science.

Worries about the explanatory nature of the alchemical, corpuscularian, and phlogiston theories are well documented (cf. Bensaude-Vincent and Stengers 1996; Brock 1993). Lavoisier and—to a lesser extent historically—Dalton initiated shifts in the explanatory tasks and presuppositions of the science. For instance, prior to Lavoisier many chemists ''explained'' a chemical by assigning it to a type associated with its experimental dispositions (e.g., flammability, acidity, etc.). After Lavoisier and Dalton, ''explanation'' most often meant the isolation and identification of a chemical's constituents. Eventually, the goal of explanation changed to the identification of the transformation processes in the reactants that gave rise to the observed properties of the intermediaries and the products. It is at that time that chemists began to write the now familiar reaction equations, which encapsulate this change. These transformations cannot be described simply as the coming of new theories; they also involved changes in explanatory presuppositions and languages, as well as new experimental techniques (Bensaude-Vincent and Stengers 1996; Nye 1993).

Thinking in terms of transformation processes led chemists to postulate atoms as the agents of the transformation process. But atoms were not observable in the nineteenth century. How is the explanatory power of these unobservable entities accounted for? Also, do chemical elements retain their identity in compounds (Paneth 1962)? Something remains the same, yet the properties that identify elements (e.g., the green color of chlorine gas) do not exist in compounds (e.g., sodium chloride, or common table salt).

The history of chemistry also raises a number of interesting questions about the character of knowledge and understanding in the science. Whereas philosophers have historically identified theoretical knowledge with laws or sets of propositions, the history of chemistry shows that there are different kinds of knowledge that function as a base for understanding. Much of chemistry is experimental, and much of what is known to be true arises in experimental practice independently of or only indirectly informed by theoretical knowledge. When chemists have theorized, they have done so freely, using and combining phenomenological, constructive (in which the values of certain variables are given by experiment or other theory), and deductive methods. Chemists have rarely been able to achieve anything like a strict set of axioms or first principles that order the phenomena and serve as their explanatory base (Bensaude-Vincent and Stengers 1996; Gavroglu 1997; Nye 1993).

Historical research has also raised the issue of whether theory has contributed most to the progress of chemistry. While philosophers often point to the conceptual ''revolution'' wrought by Lavoisier as an example of progress, historians more often point to the ways in which laboratory techniques have been an important motor of change in chemistry, by themselves or in tandem with theoretical shifts (Bensaude-Vincent and Stengers 1996; Nye 1993). For example, during the nineteenth century, substitution studies, in which one element is replaced by another in a compound, were driven largely by experimental practices. Theoretical concepts did not, in the first instance, organize the investigation (Klein 1999). Similar remarks can be made regarding the coming of modern experimental techniques such as various types of chromatography and spectroscopy (Baird 2000; Slater 2002). Chemists often characterize a molecule using such techniques, and while the techniques are grounded in physical theory, the results must often be interpreted in chemical language. These examples lead back to questions about the nature of chemical knowledge. Arguably, the knowledge appears to be a mix of ''knowing how'' and ''knowing that'' which is not based solely in the theory (chemical or physical) available at the time.

A number of the philosophical themes raised in the history of chemistry continue to reverberate in current chemistry. For instance, what is the proper ontological base for chemical theory, explanation, and practice? While many chemists would unfailingly resort to molecular structure as the explanation of what is seen while a reaction is taking place, this conception can be challenged from two

directions. From one side, echoing the eighteenth-century conception of the science, chemistry begins in the first instance with conceptions and analyses of the qualitative properties of material stuff (Schummer 1996; van Brakel 2001). One might call the ontology associated with this conception a metaphysically nonreductive dispositional realism, since it focuses on properties and how they appear under certain conditions and does not attempt to interpret them in any simpler terms. In this conception, reference to the underlying molecular structure is subsidiary or even otiose, since the focus is on the observable properties of the materials. Given that molecular structure is difficult to justify within quantum mechanics (see below), one can argue that it is justifiable to remain with the observable properties. If this view is adopted, however, the justification of the ontology of material stuff becomes a pressing matter. Quantum mechanics will not supply the justification, since it does not deliver the qualitative properties of materials. Further, it still seems necessary to account for the phenomenal success that molecular explanations afford in planning and interpreting chemical structures and reactions.

From the other side, pure quantum mechanics makes it difficult to speak of the traditional atoms-within-a-molecule approach referred to in the reaction equations and structural diagrams (Primas 1983; Weininger 1984). Quantum mechanics tells us that the interior parts of molecules should not be distinguishable; there exists only a distribution of nuclear and electronic charges. Yet chemists rely on the existence and persistence of atoms and molecules in a number of ways. To justify these practices, some have argued that one can forgo the notion of an atom based on the orbital model and instead identify spatial regions within a molecule bounded by surfaces that have a zero flux of energy across the surfaces (Bader 1990). Currently, it is an open question whether this representation falls naturally out of quantum mechanics, and so allows one to recover atoms as naturally occurring substituents of molecules, or whether the notion of 'atom' must be presupposed in order for the identification to be made. Here again there is a question of the character of the theory that will give the desired explanation.

A number of examples supporting the claim that quantum mechanics and chemistry are uneasy bedfellows will be discussed below as they relate to the issue of reductionism, but their relationship also raises forcefully the long-standing issue of how theory guides chemical practice. Even in this era of supercomputers, only the energy states of systems with relatively few electrons or with high degrees of symmetry can be calculated with a high degree of faithfulness to the complete theoretical description. For most chemical systems, various semi-empirical methods must be used to get theoretically guided results. More often than not, it is the experimental practice independent of any theoretical calculation that gets the result. Strictly theoretical predictions of novel properties are rather rare, and whether they are strictly theoretical is a matter that can be disputed. A case in point is the structure of the CH_2 molecule. Theoretical chemists claimed to have predicted novel properties of the molecule, *viz.*, its nonlinear geometry, prior to any spectroscopic evidence (Foster and Boys 1960). While it is true that the spectroscopic evidence was not yet available, it may have been the case that reference to analogous molecules allowed the researchers to set the values for some of the parameters in the equations. So the derivation may not have been as *a priori* as it seemed.

Like the other special sciences, chemistry raises the issue of reductionism quite forcefully. However, perhaps because most philosophers have accepted at face value Dirac's famous dictum that chemistry has become nothing more than the application of quantum mechanics to chemical problems (cf. Nye 1993, 248), few seem to be aware of the difficulties of making good on that claim using the tools and concepts available in traditional philosophical analyses of the sciences. No one doubts that chemical forces are physical in nature, but connecting the chemical and physical mathematical structures and/or concepts proves to be quite a challenge. Although problems involving the relation between the physical and the chemical surround a wide variety of chemical concepts, such as aromaticity, acidity (and basicity), functional groups, and substituent effects (Hoffmann 1995), three examples will be discussed here to illustrate the difficulties: the periodic table, the use of orbitals to explain bonding, and the concept of molecular shape. Each also raises issues of explanation, representation, and realism.

Philosophers and scientists commonly believe that the periodic table has been explained by—and thus reduced to—quantum mechanics. This is taken to be an explanation of the configuration of the electrons in the atom, and, as a result of this, an explanation of the periodicity of the table. In the first case, however, configurations of electrons in atoms and molecules are the result of a particular approximation, in which the many-electron quantum wavefunction is rewritten as a series of one-electron functions. In practice, these one-electron

functions are derived from the hydrogen wavefunction and, upon integration, lead to the familiar spherical *s* orbital, the dumbbell-shaped *p* orbitals, and the more complicated *d* and *f* orbitals. Via the Pauli exclusion principle, which states that the spins are to be paired if two electrons are to occupy one orbital and no more than two electrons may occupy an orbital, electrons are assigned to these orbitals. If the approximation is not made (and quantum mechanics tells us it should not be, since the approximation relies on the distinguishability of electrons), the notion of individual quantum numbers—and thus configurations—is no longer meaningful (Richman 1999). In addition, configurations themselves are not observable; absorption and emission spectra are observed and interpreted as energy transitions between orbitals of different energies.

In the second case, quantum mechanics explains only part of the periodic table, and often it does not explain the features of the table that are of most interest to chemists (Scerri 1998). Pauli's introduction of the fourth quantum number, ''spin'' or ''spin angular momentum,'' leads directly to the *Aufbau* principle, which states that the periodic table is constructed by placing electrons in lower energy levels first and then demonstrating that atoms with similar configurations have similar chemical properties. In this way, one can say that since chlorine and fluorine need one more electron to achieve a closed shell, they will behave similarly. However, this simple, unqualified explanation suffers from a number of anomalies. First, the filling sequence is not always strictly obeyed. Cobalt, nickel, and copper fill their shells in the sequence $3d^7 4s^2$, $3d^8 4s^2$, $3d^{10} 4s^1$. The superscripts denote the number of electrons in the subshell; the *s* shell can hold a maximum of two electrons and the five *d* orbitals ten. The observed order of filling is curious from the perspective of the unmodified *Aufbau* principle for a number of reasons. The 4*s* shell, which is supposed to be higher in energy than the 3*d* shell, has been occupied and closed first. Then, there is the ''demotion'' of one 4*s* electron in nickel to a 3*d* electron in copper. Second, configurations are supposed to explain why elements falling into the same group behave similarly, as in the example of chlorine and fluorine. Yet nickel, palladium, and platinum are grouped together because of their marked chemical similarities despite the fact that their outer shells have different configurations ($4s^2$, $5s^0$ and $6s^1$, respectively). These and other anomalies can be resolved using alternative derivations more closely tied to fundamental quantum mechanics, but the derivations require that the orbital approximation be dropped, and it was that approximation that was the basis for the assignment into the *s*, *p*, and *d* orbitals in the first place. It thus becomes an open question whether quantum mechanics, via the *Aufbau* principle, has explained the chemical periodicities encapsulated in the table.

Similar questions about the tenuous relation between physics and chemistry surround the concept of bonding. At a broad level, chemists employ two seemingly inconsistent representations, the valence bond (VB) and molecular orbital (MO) theories, to explain why atoms and molecules react. Both are calculational approximations inherited from atomic physics. The relations between the two theories and respective relations to the underlying quantum mechanics raise many issues of theory interpretation and realism about the chemical concepts (see below). Subsidiary concepts such as resonance are also invoked to explain the finer points of bonding. Chemists have offered competing realist interpretations of this concept, and philosophers have offered various realist and instrumentalist interpretations of it as well (Mosini 2000).

More specifically, a host of philosophical issues are raised within the molecular orbital theory. Here, bonding is pictured as due to the interaction of electrons in various orbitals. As noted earlier, the familiar spherical and dumbbell shapes arise only because of the orbital approximation. However, there is no reason to expect that the hydrogenic wavefunctions will look anything like the molecular ones (Bader 1990; Woody 2000). After the hydrogenic wavefunctions have been chosen as the basis for the calculation, they must be processed mathematically to arrive at a value for the energy of the orbital that is at all close to the experimentally observed value. The molecular wave equation is solved by taking linear combinations of the hydrogenic wavefunctions, forming the product of these combinations (the ''configuration interaction'' approach), and using the variational method to produce a minimal energy solution to the equation. The familiar orbitals appear only when these three steps in the complete solution have been omitted (Woody 2000). Thus, the idea that the familiar orbitals are responsible for the bonding is thrown into question. Yet the orbitals classify and explain how atoms and molecules bond extremely well. That they do provide deep, unified, and fertile representations and explanations seems curious from the perspective of fundamental quantum mechanics. Clearly, more analysis is required to understand the relation clearly. If one insists on a philosophical account of reduction that requires the mathematical or logical derivability of one theory from another, how orbitals achieve their power

remains obscure. Even when one abandons that philosophical account, it is not clear how the representations have the organizing and explanatory power they do (see Reductionism).

As a final illustration of the difficulty of connecting physics and chemistry in any sort of strict fashion, consider the concept of molecular shape. Partly through the tradition of orbitals described above and partly through a historical tradition of oriented bonding that arose well before the concept of orbitals was introduced, chemists commonly explain many behaviors of molecules as due to their three-dimensional orientiation in space. Molecules clearly react as if they are oriented in three-dimensional space. For example, the reaction $I^- + CH_3Br \rightarrow ICH_3 + Br^-$ is readily explained by invoking the notion that the iodine ion (I^-) attacks the carbon (C) on the side away from the bromine (Br) atom. (This can be detected by substituting deuterium atoms for one of the hydrogens [H] and measuring subsequent changes in spectroscopic properties.) As noted before, however, such explanations are suspect within quantum mechanics, since talk of oriented bonds and quasi-independent substituents in the reaction is questionable. Orientations must be ''built into'' the theory by parameterizing some of the theoretical variables. Unfortunately, there is no strict quantum mechanical justification for the method by which orientation in space is derived. Orientation relies on a notion of a nuclear frame surrounded by electrons. This notion is constructed via the Born-Oppenheimer approximation, which provides a physical rationale for why the nuclear positions should be slowly varying with respect to the electronic motions. In some measurement regimes, the approximation is invalid, and correct predictions are achieved only by resorting to a more general molecular Hamiltonian. In more common measurement regimes, the approximation is clearly valid. But, as with the issues involved in the case of the periodic table and orbitals, the physics alone do not tell us why it is valid. There is a physical justification for the procedure, but this justification has no natural representation with the available physical theory (Weininger 1984). Should justification based on past experience be trusted, or should the theory correct the interpretive practice? In any case, the chemistry is consistent with, but not yet derivable from, the physics.

All three examples are connected with a methodological issue mentioned earlier, *viz*., the type of theory that chemists find useful. As previously noted, chemists often must parameterize the physical theories at their disposal to make them useful. All three of the cases described above involve such parameterization, albeit in different ways. How is it that such parameterizations uncover useful patterns in the data? Are they explanatory? When are they acceptable and when not (Ramsey 1997)? These and a host of similar questions remain to be answered.

Other epistemological and ontological issues raised in the practice of chemistry remain virtually unexplored. For instance, the question of whether one molecule is identical to another is answered by referring to some set of properties shared by the two samples. Yet the classification of two molecules as of the ''same'' type will vary, since different theoretical representations and experimental techniques detect quite different properties (Hoffmann 1995). For instance, reference can be made to the space-filling property of molecules, their three-dimensional structure, or the way they respond to an electric field. Additionally, the determination of sameness must be made in light of the question, For what function or purpose? For instance, two molecules of hemoglobin, which are large biological molecules, might have different isotopes of oxygen at one position. While this difference might be useful in order to discover the detailed structure of the hemoglobin molecule, it is usually irrelevant when talking about the molecule's biological function.

How the explanatory practices of chemistry stand in relation to the available philosophical accounts and to the practices of other sciences remains an important question. Chemical explanations are very specific, often lacking the generality invoked in philosophical accounts of explanation. Moreover, chemists invoke a wide variety of models, laws, theories, and mechanisms to explain the behavior and structure of molecules. Finally, most explanations require analogically based and/or experimentally derived adjustments to the theoretical laws and regularities in order for the account to be explanatory.

As mentioned earlier, chemistry is an extremely experimental science. In addition to unifying and fragmenting research programs in chemistry, new laboratory techniques have dramatically changed the epistemology of detection and observation in chemistry (e.g., from tapping manometers to reading NMR [nuclear magnetic resonance] outputs). As yet, however, there is no overarching, complete study of the changes in the epistemology of experimention in chemistry: for example, what chemists count as observable (and how this is connected to what they consider to be real), what they assume counts as a complete explanation, what they assume counts as a successful end to an experiment, etc.

Additionally, what are the relations between academic and industrial chemistry? What are the relative roles of skill, theory, and experiment in

these two arenas of inquiry? Last but not least, there are pressing ethical questions. The world has been transformed by chemical products. Chemistry has blurred the distinction between the natural and the artificial in confusing ways (Hoffmann 1995). For instance, catalytically produced ethanol is chemically identical with the ethanol produced in fermentation. So is the carbon dioxide produced in a forest fire and in a car's exhaust. Why are there worries about the exhaust fumes but not the industrially produced ethanol? Is this the appropriate attitude? Last but not least, chemicals have often replaced earlier dangerous substances and practices; witness the great number of herbicides and insecticides available at the local garden center and the prescription medicines available at the pharmacy. Yet these replacements are often associated with a cost. One need think only of DDT or thalidomide to be flung headlong into ethical questions regarding the harmfulness and use of human-made products.

Many of the above topics have not been analyzed in any great depth. Much remains to done to explore the methodology and philosophy of the chemical sciences.

JEFFRY L. RAMSEY

AU:21

References

Bader, R. (1990), Atoms in Molecules: a Quantum Theory. New York: Oxford University Press.

Baird, D. (2000), "Encapsulating Knowledge: The Direct Reading Spectrometer," *Foundations of Chemistry* 2: 5–46.

Bensaude-Vincent, B., and I. Stengers (1996), *A History of Chemistry*. Translated by D. van Dam. Cambridge, MA: Harvard University Press.

Bhushan, N., and S. Rosenfeld (2000), *Of Minds and Molecules: New Philosophical Perspectives on Chemistry*. New York: Oxford University Press.

Brock, W. (1993), *The Norton History of Chemistry*. New York: W. W. Norton.

Foster, J. M., and S. F. Boys (1960), "Quantum Variational Calculations for a Range of CH_2 Configurations," *Reviews of Modern Physics* 32: 305–307.

Gavroglu, K. (1997), "Philosophical Issues in the History of Chemistry," *Synthese* 111: 283–304.

Hoffmann, R. (1995), *The Same and Not the Same*. New York: Columbia University Press.

Klein, U. (1999), "Techniques of Modeling and Paper-Tools in Classical Chemistry," in M. Morgan and M. Morrison (eds.), *Models as Mediators*. New York: Cambridge University Press, 146–167.

Mosini, V. (2000), "A Brief History of the Theory of Resonance and of Its Interpretation," *Studies in History and Philosophy of Modern Physics* 31B: 569–581.

Nye, M. J. (1993), *From Chemical Philosophy to Theoretical Chemistry*. Berkeley and Los Angeles: University of California Press.

Paneth, F. (1962), "The Epistemological Status of the Concept of Element," *British Journal for the Philosophy of Science* 13: 1–14, 144–160.

Primas, H. (1983), *Chemistry, Quantum Mechanics and Reductionism*. New York: Springer Verlag.

Ramsey, J. (1997), "Between the Fundamental and the Phenomenological: The Challenge of 'Semi-Empirical' Methods," *Philosophy of Science* 64: 627–653.

Scerri, E. (1998), "How Good Is the Quantum Mechanical Explanation of the Periodic System?" *Journal of Chemical Education* 75: 1384–1385.

Schummer, J. (1996), *Realismus und Chemie*. Wuerzburg: Koenigshausen and Neumann.

Slater, L. (2002), "Instruments and Rules: R. B. Woodward and the Tools of Twentieth-century Organic Chemistry," *Studies in History and Philosophy of Science* 33: 1–32.

van Brakel, J. (2001), *Philosophy of Chemistry: Between the Manifest and the Scientific Image*. Leuven, Belgium: Leuven University Press.

Weininger, S. (1984), "The Molecular Structure Conundrum: Can Classical Chemistry Be Reduced to Quantum Chemistry?" *Journal of Chemical Education* 61: 939–944.

Woody, A. (2000), "Putting Quantum Mechanics to Work in Chemistry: The Power of Diagrammatic Representation," *Philosophy of Science* 67(suppl): S612–S627.

See also **Explanation; Laws of Nature; Quantum Mechanics; Reductionism**

CHOMSKY, NOAM

(7 December 1928–)

Avram Noam Chomsky received his Ph.D in linguistics from the University of Pennsylvania and has been teaching at Massachusetts Institute of Technology since 1955, where he is currently Institute Professor. Philosophers are often familiar with the early work of Chomsky (1956, 1957, 1959a,

and 1965), which applied the methods of formal language theory to empirical linguistics, but his work has also incorporated a number of philosophical assumptions about the nature of scientific practice—many of which are defended in his writings.

This essay will first describe the development and evolution of Chomsky's theory of generative linguistics, highlighting some of the philosophical assumptions that have been in play. It will then turn to some of the methodological debates in generative linguistics (and scientific practice more generally), focusing on Chomsky's role in these debates.

The Development and Evolution of Generative Grammar

A number of commentators have suggested that Chomsky's early work in generative linguistics initiated a kind of Kuhnian paradigm shift in linguistic theory. While Chomsky himself would reject this characterization (at least for his initial work in generative grammar), it is instructive to examine the development of generative linguistics, for it provides an excellent laboratory for the study of the development of a young science, and in particular it illuminates some of the philosophical prejudice that a young science is bound to encounter.

Chomsky's role in the development of linguistic theory and cognitive science generally can best be appreciated if his work is placed in the context of the prevailing intellectual climate in the 1950s—one in which behaviorism held sway in psychology departments and a doctrine known as American Structuralism was prevalent in linguistics departments.

American Structuralism, in particular as articulated by Bloomfield (1933 and 1939), adopted a number of key assumptions that were in turn adopted from logical empiricism (see Logical Empiricism). Newmeyer (1986, ch. 1) notes that the following assumptions were in play:

- All useful generalizations are inductive generalizations.
- Meanings are to be eschewed because they are occult entities—that is, because they are not directly empirically observable.
- Discovery procedures like those advocated in logical empiricism should be developed for the proper conduct of linguistic inquiry.
- There should be no unobserved processes.

One of the ways in which these assumptions translated into theory was in the order that various levels of linguistic description were to be tackled. The American Structuralists identified four levels: phonemics (intuitively the study of sound patterns), morphemics (the study of words, their prefixes and suffixes), syntax (the study of sentence-level structure), and discourse (the study of cross-sentential phenomena). The idea was that proper methodology would dictate that one begin at the level of phonemics, presumably because it is closer to the data; then proceed to construct a theory of morphemics on the foundations of phonemics; then proceed to construct a theory of syntax, etc.

Notice the role that the concepts of logical empiricism played in this proposed methodology. One finds radical reductionism in the idea that every level must be reducible to the more basic phonemic level; verificationism in the contention that the phonemic level is closely tied to sense experience; and discovery procedures in the suggestion that this overall order of inquiry should be adopted (see Reductionism; Verificationism).

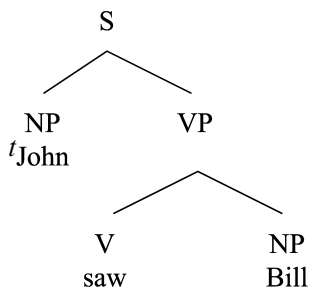
Chomsky rejected most if not all of these assumptions early on (see Chomsky [1955] 1975, introduction, for a detailed discussion). As regards discovery procedures, for example, he rejected them while still a matriculating graduate student, then holding a position in the Harvard Society of Fellows:

> By 1953, I came to the same conclusion [as Morris Halle]if the discovery procedures did not work, it was not because I had failed to formulate them correctly, but because the entire approach was wrong. ... [S]everal years of intense effort devoted to improving discovery procedures had come to naught, while work I had been doing during the same period on generative grammars and explanatory theory, in almost complete isolation, seemed to be consistently yielding interesting results. (1979: 131)

Chomsky also rejected the assumption that all processes should be "observable"—early theories of transformational grammar offered key examples of unobservable processes. For example, in his "aspects theory" of generative grammar (Chomsky 1965), the grammar is divided into two different "levels of representation," termed initially *deep structure* and *surface structure*. The deep-structure representations were generated by a context-free phase structure grammar—that is, by rules (of decomposition, "→") of the following form, where S stands for sentence, NP for noun phrase, VP for verb phrase, etc.

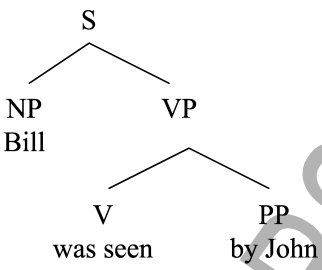
$$S \rightarrow NP\ VP$$
$$VP \rightarrow V\ NP$$
$$NP \rightarrow John$$
$$NP \rightarrow Bill$$
$$V \rightarrow saw$$

These rewriting rules then generated linguistic representations of the following form:



(1)

Crucially for Chomsky, the objects of analysis in linguistic theory were not the terminal strings of words, but rather *phrase markers*—structured objects like (1). Transformational rules then operated on these deep-structure representations to yield surface-structure representations. So, for example, the operation of passivization would take a deep-structure representation like (1) and yield the surface-structure representation (abstracting from detail) in (2):



(2)

The sentence in (3) is therefore a complex object consisting of (at a minimum) an ordered pair of the two representations corresponding to (1) and (2):

Bill was seen by John. (3)

Clearly, Chomsky was committed not only to "unobserved" processes in the guise of transformations, but also to unobserved levels of representation. No less significant was the nature of the data that Chomsky admitted—not utterances or written strings, but rather speakers' judgments of acceptability and meaning. Thus, (3) is not a datum because it has been written or spoken, but rather because speakers have intuitions that it is (would be) an acceptable utterance. Here again, Chomsky broke with prevailing methodology in structuralist linguistics and, indeed, behaviorist psychology, by allowing intuitions rather than publicly available behaviors as data.

Generative grammar subsequently evolved in response to a number of internal pressures. Crucially, the number of transformations began to proliferate in a way that Chomsky found unacceptable. Why unacceptable? Early on in the development of generative grammar, Chomsky had made a distinction between the *descriptive adequacy* and the *explanatory adaquacy* of an empirical linguistic theory (Chomsky 1965 and 1986b). In particular, if a linguistic theory is to be explanatorially adequate, it must not merely describe the facts, but must do so in a way that explains how humans are able to learn languages. Thus, linguistics was supposed to be embeddable into cognitive science more broadly. But if this is the case, then there is a concern about the unchecked proliferation of rules—such rule systems might be descriptively adequate, but they would fail to account for how we learn languages (perhaps due to the burden of having to learn all those language-specific rules).

Chomsky's initial (1964 and 1965) solution to this problem involved the introduction of conditions on transformations (or constraints on movement), with the goal of reducing the complexity of the descriptive grammar. In Chomsky (1965), for example, the recursive power of the grammar is shifted from the transformations to the phrase structure rules alone. In the "extended standard theory" of the 1970s, there was a reduction of the phrase structure component with the introduction of "X-bar theory," and a simplification of the constraints on movement. This was followed by a number of proposals to reduce the number and types of movement rules themselves. This came to a head (Chomsky 1977 and 1981a) with the abandonment of specific transformations altogether for a single rule ("move-α"), which stated, in effect, that one could move anything anywhere. This one rule was then supplemented with a number of constraints on movement. As Chomsky and a number of other generative linguists were able to show, it was possible to reduce a great number of transformations to a single-rule move-α and to a handful of constraints on movement.

Chomsky (1981b, 1982, and 1986a) synthesized subsequent work undertaken by linguists working in a number of languages, ranging from the romance languages to Chinese and Japanese, showing that other natural languages had similar but not identical constraints, and it was hypothesized that the variation was due to some limited parametric variation among human languages. This work established the

"principles and parameters" framework of generative grammar. To get an idea of this framework, consider the following analogy: Think of the language faculty as a prewired box containing a number of switches. When exposed to environmental data, new switch settings are established. Applying this metaphor, the task of the linguist is to study the initial state of the language faculty, determine the possible parametric variations (switch settings), and account for language variation in terms of a limited range of variation in parameter settings. In Chomsky's view (2000, 8), the principles-and-parameters framework "gives at least an outline of a genuine theory of language, really for the first time." Commentators (e.g., Smith 2000, p. xi) have gone so far as to say that it is "the first really novel approach to language of the last two and a half thousand years." In what sense is it a radical departure? For the first time it allowed linguists to get away from simply constructing rule systems for individual languages and to begin exploring in a deep way the underlying similarities of human languages (even across different language families), to illuminate the principles that account for those similarities, and ultimately to show how those principles are grounded in the mind/brain. In Chomsky's view, the principles-and-parameters framework has yielded a number of promising results, ranging from the discovery of important similarities between *prima facie* radically different languages like Chinese and English to insights into the related studies of language acquisition, language processing, and acquired linguistic deficits (e.g., aphasia). Perhaps most importantly, the principles-and-parameters framework offered a way to resolve the tension between the two goals of descriptive adequacy and explanatory adequacy.

Still working within the general principles-and-parameters framework, Chomsky (1995 and 2000, ch. 1) has recently articulated a research program that has come to be known as the "minimalist program," the main idea behind which is the working hypothesis that the language faculty is not the product of messy evolutionary tinkering—for example, there is no redundancy, and the only resources at work are those that are driven by "conceptual necessity." Chomsky (1995, ch. 1) initially seemed to hold that in this respect the language faculty would be unlike other biological functions, but more recently (2001) he seems to be drawn to D'Arcy Thompson's theory that the core of evolutionary theory consists of physical/mathematical/chemical principles that sharply constrain the possible range of organisms. In this case, the idea would be not only that those principles constrain low-level biological processes (like sphere packing in cell division) but also that such factors might be involved across the board—even including the human brain and its language faculty.

In broadest outline, the minimalist program works as follows: There are two levels of linguistic representation, phonetic form (PF) and logical form (LF), and a well-formed sentence (or linguistic *structure*) must be an ordered pair $<\pi, \lambda>$ of these representations (where π is a phonetic form and λ is the logical form). PF is taken to be the level of representation that is the input to the performance system (e.g., speech generation), and LF is, in Chomsky's terminology, the input to the conceptual/intensional system. Since language is, if nothing else, involved with the pairing of sounds and meanings, these two levels of representation are conceptually necessary. A minimal theory would posit no other levels of representation.

It is assumed that each sentence (or better, *structure*, Σ) is constructed out of an *array* or *numeration*, N, of lexical items. Some of the items in the numeration will be part of the pronounced (written) sentence, and others will be part of a universal inventory of lexical items freely inserted into all numerations. Given the numeration N, the computational system (C_{HL}) attempts to derive (compute) well-formed PF and LF representations, converging on the pair $<\pi, \lambda>$. The derivation is said to *converge* at a certain level if it yields a representation that is interpretable at that level. If it fails to yield an interpretable representation, the derivation *crashes*. Not all converging derivations yield structures that belong to a given language L. Derivations must also meet certain economy conditions.

Chomsky (2000, 9) notes that the import of the minimalist program is not yet clear. As matters currently stand, it is a subresearch program within the principles-and-parameters framework that is showing some signs of progress—at least enough to encourage those working within the program. As always, the concerns are to keep the number of principles constrained, not just to satisfy economy constraints, but to better facilitate the embedding of linguistics into theories of language acquisition, cognitive psychology, and, perhaps most importantly, general biology.

Some Conceptual Issues in Generative Grammar

While Chomsky would argue that he does not have a philosophy of science per se and that his

philosophical observations largely amount to common sense, a number of interesting debates have arisen in the wake of his work. The remainder of this essay will review some of those debates.

On the Object of Study

AU:22 Chomsky (1986) draws the distinction between the notions of *I*-language and *E*-language, where *I*-language is the language faculty discussed above, construed as a chapter in cognitive psychology and ultimately human biology. *E*-language, on the other hand, comprises a loose collection of theories that take language to be a shared social object, established by convention and developed for purposes of communication; or an abstract mathematical object of some sort.

In Chomsky's view (widely shared by linguists), the notion of a 'language' as it is ordinarily construed by philosophers of language is fundamentally incoherent. We may talk about "the English language" or "the French language" but these are loose ways of talking. Typically, the question of who counts as speaking a particular language is determined more by political boundaries than actual linguistic variation. For example, there are dialects of German that, from a linguistic point of view, are closer to Dutch than to standard German. Likewise, in the Italian linguistic situation, there are a number of so-called dialects only some of which are recognized as "official" languages by the Italian government. Are the official languages intrinsically different from the "mere" dialects? Not in any linguistic sense. The decision to recognize the former as official is entirely a political decision. In the words attributed to Max Weinreich: A language is a dialect with an army and a navy. In this case, a language is a dialect with substantial political clout and maybe a threat of separatism.

Chomsky (1994 and 2000, ch. 2) compares talk of languages (i.e., *E*-languages) to saying that two cities are "near" each other; whether two cities are near depends on our interests and our mode of transportation and very little on brute facts of geography. In the study of language, the notion of 'sameness' is no more respectable than that of 'nearness'in geography. Informally we might group together ways of speaking that seem to be similar (relative to our interests), but such groupings have no real scientific merit. As a subject of natural inquiry, the key object of study has to be the language faculty and its set of possible parametric variations.

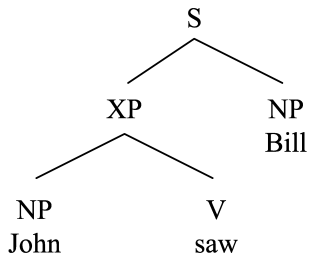
Not only is the notion of an *E*-language problematic, but it will not help to retreat to talk about *E-dialects*. The problem is that what counts as a separate *E*-dialect is also incoherent from a scientific point of view. For example, Chomsky (2000, 27) reports that in his idiolect, the word "ladder" rhymes with "matter" but not with "madder." For others, the facts do not cut in this way. Do they speak the same dialect as Chomsky or not? There is no empirical fact of the matter here; it all depends on individuals' desires to identify linguistically with each other. Even appeals to mutual intelligibility will not do, since what we count as intelligible will depend much more on our patience, our ambition, and our familiarity with the practices of our interlocutors than it will on brute linguistic facts.

If the notion of *E*-language and *E*-dialect are incoherent, is it possible to construct a notion of *E-idiolect*?—that is, to identify idiolects by external criteria like an individual's spoken or written language? Apparently not. Included in what a person says or writes are numerous slips of the tongue, performance errors, etc. How are we to rule those out of the individual's *E*-idiolect? Appeal to the agent's linguistic community will not do, since that would in turn require appeal to an *E*-language, and for the reasons outlined above, there is no meaningful way to individuate *E*-languages. In the *I*-language approach, however, the problem of individuating *I*-idiolects takes the form of a coherent empirical research project. The idiolect (*I*-idiolect) is determined by the parametric state of *A*'s language faculty, and the language faculty thus determines *A*'s linguistic *competence*. Speech production that diverges from this competence can be attributed to *performance* errors. Thus, the competence/performance distinction is introduced to illuminate the distinction between sounds and interpretations that are part of *A*'s grammar and those that are simply mistakes. The E-language perspective has no similar recourse.

The Underdetermination of Theory by Evidence

One of the first philosophical issues to fall out from the development of generative grammar has been the dispute between Quine (1972) and AU:23 Chomsky (1969 and 2000, ch. 3) on the indeterminacy of grammar. Similar to his argument for the indeterminacy of meaning, Quine held that there is no way to adjudicate between two descriptively adequate sets of grammatical rules. So, for example, imagine two rule sets, one envisioning

structures like (1) above, and another positing the following:



(4)

Chomsky maintained that Quine's argument is simply a recapitulation of the standard scientific problem of the underdetermination of theory by evidence. And in any case it is not clear that there are no linguistic tests that would allow us to choose between (1) and (4)—there are, after all, "constituency tests" (e.g., involving movement possibilities) that would allow us to determine whether the XP or the VP is the more plausible constituent.

Even if there are several grammars that are consistent with the available linguistic facts (the facts are not linguistic *behavior*, for Chomsky, but intuitions about acceptability and possible interpretation), we still have the additional constraint of which theory best accounts for the problem of language acquisition, acquired linguistic deficits (e.g., from brain damage), linguistic processing, etc. In other words, since grammatical theory is embedded within cognitive psychology, the choice between candidate theories can, in principle, be radically constrained. But further, even if there were two descriptively adequate grammars, each of which could be naturally embedded within cognitive psychology, there remain standard best-theory criteria (simplicity, etc.) that can help us to adjudicate between the theories.

Intrinsic Versus Relational Properties in Linguistics

In the philosophy of science, it is routine to make the distinction between "intrinsic" and "relational" properties. So, for example, the rest mass of an object would be an intrinsic property and its weight would be a relational property (since it depends upon the mass of the body that the object is standing on). Similarly, in the philosophy of psychology there is a common distinction between "individualistic" and "externalist" properties. So, for example, there is the question of whether psychological states supervene on individualistic properties (intrinsic properties that hold of the individual in isolation) or whether they supervene on "externalist" properties (in effect, on relations between the agent and its environment) (see Supervenience). Chomsky has argued that all scientifically interesting psychological and linguistic properties supervene upon individualist (intrinsic) properties. In particular, there is a brute fact about the state of an individual's language faculty, and that fact is determined in turn by facts about the individual in isolation—not by the environment in which the individual is embedded. Because the language faculty is part of our biological endowment, the nature of the representations utilized by the language faculty are fixed by biology and are not sensitive to environmental issues such as whether one is moving about on Earth or a phenomenologically identical planet with a different microstructure (e.g., Putnam's "Twin Earth").

Thus Chomsky takes issue with philosophers like Burge (1986), who argues in *Individualism and Psychology* that the content of the representations posited in psychology are determined at least in part by environmental factors. Chomsky holds that if the notion of content involves externalist or environmental notions, then it is not clear that it can play an interesting role in naturalistic inquiry in cognitive psychology (see Psychology, Philosophy of).

If environmentalism is to be rejected in psychology, then it naturally must be rejected in the semantics of natural language as well. That is: if the task of the linguist is to investigate the nature of *I*-language; if the nature of *I*-language is a chapter of cognitive psychology; and if cognitive psychology is an individualistic rather than a relational science, semantics will want to eschew relational properties like reference (where 'reference' is construed as a relation between a linguistic form and some object in the external environment). Thus Chomsky (1975 and 2000) rejects the notion of reference that has been central to the philosophy of language for the past three decades, characterizing it as an ill-defined technical notion (certainly one with no empirical applications), and, following Strawson, suggests that in the informal usage, individuals 'refer' but linguistic objects do not (see Reference, Theories of).

This conclusion has immediate results for the notion of theory change in science. If the notion of reference is suspect or incoherent, then it can hardly be employed in an account of theory change (as in Putnam 1988). How then to make sense of theory change? Chomsky (2000) suggests the following:

> Some of the motivation for externalist approaches derives from the concern to make sense of the history of science. Thus, Putnam argues that we should take the early Niels Bohr to have been referring to electrons in the quantum-theoretic sense, or we would have to "dismiss all of his 1900 beliefs as totally wrong" (Putnam

> 1988), perhaps on a par with someone's beliefs about angels, a conclusion that is plainly absurd . . . .

Agreeing . . . that an interest in intelligibility in scientific discourse across time is a fair enough concern, still it cannot serve as the basis for a general theory of meaning; it is, after all, only one concern among many, and not a central one for the study of human psychology. Furthermore, there are internalist paraphrases. Thus we might say that in Bohr's earlier usage, he expressed beliefs that were literally false, because there was nothing of the sort he had in mind in referring to electrons; but his picture of the world and articulation of it was structurally similar enough to later conceptions so that we can distinguish his beliefs about electrons from beliefs about angels. (43) (see Putnam, Hilary).

Against Teleological Explanation in Linguistics

A number of philosophers and linguists have thought that some progress can be made in the understanding of language by thinking of it as principally being a medium of communication. Chomsky rejects this conception of the nature of language, arguing that the language faculty is not *for* communication in any interesting sense. Of course, by rejecting the contention that language is a social object established for purposes of communication, Chomsky has not left much room for thinking of language in this way. But he also rejects standard claims that *I*-language must have evolved for the selectional value of communication; he regards such claims as without basis in fact. On this score, Chomsky sides with Gould (1980), Lewontin (1990), and many evolutionary biologists in supposing that many of our features (cognitive or anatomical) did not necessarily evolve for selectional reasons, but may have been the result of arbitrary hereditary changes that have perhaps been co-opted (see Evolution). Thus the language faculty may not have evolved for purposes of communication but may have been co-opted for that purpose, despite its nonoptimal design for communicative purposes.

In any case, Chomsky cautions that even if there was selectional pressure for the language faculty to serve as a means of communication, selection is but one of many factors in the emerging system. Crucially (2000, 163), "physical law provides narrow channels within which complex organisms may vary," and natural selection is only one factor that determines how creatures may vary within these constraints. Other factors (as Darwin himself noted) will include nonadaptive modifications and unselected functions that are determined from structure (see Function).

Inductive Versus Abductive Learning

A number of debates have turned on whether language acquisition requires a dedicated language faculty or whether "general intelligence" is enough to account for our linguistic competence. Chomsky considers the general-intelligence thesis hopelessly vague, and argues that generalized inductive learning mechanisms make the wrong predictions about which hypotheses children would select in a number of cases. Consider the following two examples from Chomsky (1975 and [1980] 1992a):

The man is tall (5)

Is the man tall? (6)

Chomsky observes that confronted with evidence of question formation like that (5)–(6) and given a choice between hypothesis (H1) and (H2), the generalized inductive learning mechanism will select (H1):

Move the first "is" to the front of the sentence. (H1)

Move the first "is" following the first NP to the front of the sentence. (H2)

But children apparently select (H2), since in forming a question from (7), they never make the error of producing (8), but always opt for (9):

The man who is here is tall. (7)

*Is the man who here is tall? (8)

Is the man who is here tall? (9)

Note that this is true despite the fact that the only data they have been confronted with previous to encountering (7) is simple data like (5)–(6). Chomsky's conclusion is that whatever accounts for children's acquisition of language, it cannot be generalized inductive learning mechanisms, but rather must be a system with structure-dependent principles/rules. Chomsky (1975, ch.1; 2000, 80) compares such learning to Peircian abduction. In other contexts this has been cast as a thesis about the "modularity" of language—that is, that there is a dedicated language acquisition device, and it, rather than some vague notion of general intelligence, accounts for our acquisition of language (see Evolutionary Psychology).

Putnam (1992) once characterized Chomsky's notion of a mental organ like the language faculty as being "biologically inexplicable," but Chomsky ([1980] 1992b) has held that it is merely "unexplained" and not inexplicable; in his view the language faculty is thus on the same footing as many other features of our biology (see Abduction; Inductive Logic).

The Science-Forming Faculty

On a more general and perhaps more abstract level, Chomsky has often spoken of a "science-forming faculty," parallel to our language faculty. The idea is that science could not have formed in response to mere inductive generalizations, but that human beings have an innate capacity to develop scientific theories (Chomsky credits C. S. Peirce with this basic idea). Despite Star Trekkian assumptions to the contrary, extraterrestrials presumably do not have a science-forming faculty like we do, and may go about theorizing in entirely different ways, which would not be recognized as "scientific" by humans.

While the science-forming faculty may be limited, Chomsky would not concede that its limits are necessarily imposed by the selectional pressures of our prehistoric past. Just as the language faculty may not have evolved principally in response to selectional pressures, so too the science-forming faculty may have emerged quite independently of selectional considerations. As Chomsky (2000, ch. 6) notes (citing Lewontin 1990), insects may seem marvelously well adapted to flowering plants, but in fact insects evolved to almost their current diversity and structure millions of years before flowering plants existed. Perhaps it is a similar situation with the science-forming faculty. That is, perhaps it is simply a matter of good fortune for us that the science-forming faculty is reliable and useful, since it could not have evolved to help us with quantum physics, for example.

The Limits of Science

The notion of a science-forming faculty also raises some interesting questions about the limits of our ability to understand the world. Since what we can know through naturalistic endeavors is bounded by our science-forming faculty, which in turn is part of our biological endowment, it stands to reason that there are questions that will remain mysteries to us—or at least outside the scope of naturalistic inquiry:

> Like other biological systems, SFF [the science-forming faculty] has its potential scope and limits; we may distinguish between *problems* that in principle fall within its range, and *mysteries* that do not. The distinction is relative to humans; rats and Martians have different problems and mysteries and, in the case of rats, we even know a fair amount about them. The distinction also need not be sharp, though we certainly expect it to exist, for any organism and any cognitive faculty. The successful natural sciences, then, fall within the intersection of the scope of SFF and the nature of the world; they treat the (scattered and limited) aspects of the world that we can grasp and comprehend by naturalistic inquiry, in principle. The intersection is a chance product of human nature. Contrary to speculations since Peirce, there is nothing in the theory of evolution, or any other intelligible source, that suggests that it should include answers to serious questions we raise, or even that we should be able to formulate questions properly in areas of puzzlement. (2000, ch. 4)

The question of what the "natural" sciences are, then, might be answered, narrowly, by asking what they have achieved; or more generally, by inquiry into a particular faculty of (the human) mind, with its specific properties.

The Mind/Body Problem and the Question of Physicalism

Chomsky has consistently defended a form of methodological monism (he is certainly no dualist); but for all that, he is likewise no materialist. In Chomsky's view the entire mind/body question is ill-formed, since there is no coherent notion of physical body. This latter claim is not in itself unique; Crane and Mellor (1990) have made a similar point. There is a difference, however; for Crane and Mellor, developments in twentieth-century science have undermined physicalism, but for Chomsky the notion of physical body was already undermined by the time of Newton:

> Just as the mechanical philosophy appeared to be triumphant, it was demolished by Newton, who reintroduced a kind of "occult" cause and quality, much to the dismay of leading scientists of the day, and of Newton himself. The Cartesian theory of mind (such as it was) was unaffected by his discoveries, but the theory of body was demonstrated to be untenable. To put it differently, Newton eliminated the problem of "the ghost in the machine" by exorcising the machine; the ghost was unaffected. (2000, 84)

In Chomsky's view, then, investigations into the mind (in the guise of cognitive science generally or linguistics in particular) can currently proceed without worrying about whether they hook up with what we know about the brain, or even fundamental particles. The unification of science remains a goal, but in Chomsky's view it is not

the study of mind that must be revised so as to conform to physical theory, but rather physical theory may eventually have to incorporate what we learn in our study of the mind. According to Chomsky this is parallel to the situation that held prior to the unification of chemistry and physics; it was not chemistry that needed to be modified to account for what we know about physics, but in fact just the opposite:

> Large-scale reduction is not the usual pattern; one should not be misled by such dramatic examples as the reduction of much of biology to biochemistry in the middle of the twentieth century. Repeatedly, the more ''fundamental'' science has had to be revised, sometimes radically, for unification to proceed. (2000, 82)

Conclusion

The influence of Chomsky's work has been felt in a number of sciences, but perhaps the greatest influence has been within the various branches of cognitive science. Indeed, Gardner (1987) has remarked that Chomsky has been the single most important figure in the development of cognitive science. Some of Chomsky's impact is due to his role in arguing against behaviorist philosophers such as Quine; some of it is due to work that led to the integration of linguistic theory with other sciences; some of it is due to the development of formal tools that were later employed in disciplines ranging from formal language theory (cf. the ''Chomsky Hierarchy'') to natural language processing; and some of it is due to his directly engaging psychologists on their own turf. (One classic example of this was Chomsky's [1959b] devastating review of Skinner's [1957] *Verbal Behavior*. See also his contributions to Piatelli-Palmerini 1980) (See Behaviorism).

With respect to his debates with various philosophers, Chomsky has sought to expose what he has taken to be double standards in the philosophical literature. In particular he has held that while other sciences are allowed to proceed where inquiry takes them without criticism by armchair philosophers, matters change when the domain of inquiry shifts to mind and language:

> The idea is by now a commonplace with regard to physics; it is a rare philosopher who would scoff at its weird and counterintuitive principles as contrary to right thinking and therefore untenable. But this standpoint is commonly regarded as inapplicable to cognitive science, linguistics in particular. Somewhere in-between, there is a boundary. Within that boundary, science is self-justifying; the critical analyst seeks to learn about the criteria for rationality and justification from the study of scientific success. Beyond that boundary, everything changes; the critic applies independent criteria to sit in judgment over the theories advanced and the entitities they postulate. This seems to be nothing more than a kind of ''methodological dualism,'' far more pernicious than the tradtional metaphysical dualism, which was a scientific hypothesis, naturalistic in spirit. Abandoning this dualist stance, we pursue inquiry where it leads. (2000, 112)

PETER LUDLOW

References

Bloomfield, L. (1933), *Language*. New York: Holt, Rinehart and Winston.

——— (1939), *Linguistic Aspects of Science: International Encyclopedia of Unified Science* (Vol. 1, No. 4). Chicago: University of Chicago Press.

Burge, T. (1986), ''Individualism and Psychology,'' *Philosophical Review* 95: 3–45.

Chomsky, N. ([1955] 1975), *The Logical Structure of Linguistic Theory*. Chicago: University of Chicago Press.

——— (1956), ''Three Models for the Description of Language,'' *I.R.E. Transactions of Information Theory* IT-2: 113–124.

——— (1957), *Syntactic Structures*. The Hague: Mouton.

——— (1959a), ''On Certain Formal Properties of Grammars,'' *Information and Control* 2: 137–167.

——— (1959b), ''Review of B. F. Skinner, *Verbal Behavior*,'' *Language* 35, 26–57.

——— (1964), *Current Issues in Linguistic Theory*. The Hague: Mouton & Co.

——— (1965), *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

——— (1968), *Language and Mind*. New York: Harcourt Brace Jovanovich.

——— (1969), ''Quine's Empirical Assumptions,'' in D. Davidson and J. Hintikka (eds.), *Words and Objections: Essays on the Work of W.V. Quine*. Dordrecht, Netherlands: D. Reidel.

——— (1975), *Reflections on Language*. New York: Pantheon.

——— (1977), ''Conditions on Rules of Grammar,'' in *Essays on Form and Interpretation*. Amsterdam: Elsevier North-Holland, 163–210.

——— (1979), *Language and Responsibility*. New York: Pantheon.

——— (1980), *Rules and Representations*. New York: Columbia University Press.

——— (1981a), *Lectures on Government and Binding*. Dordrecht, Netherlands: Foris Publications.

——— (1981b), ''Principles and Parameters in Syntactic Theory,'' in N. Horstein and D. Lightfoot (eds.), *Explanation in Linguistics: The Logical Problem of Language Acquisition*. London: Longman.

——— (1982), *Some Concepts and Consequences of the Theory of Government and Binding*. Cambridge, MA: MIT Press.

——— (1986a), *Barriers*. Cambridge, MA: MIT Press.

——— (1986b), *Knowledge of Language*. New York: Praeger.

——— (1988), *Language and Problems of Knowledge: The Managua Lectures*. Cambridge, MA: MIT Press.
——— ([1980] 1992a), "On Cognitive Structures and Their Development," in B. Beakley and P. Ludlow (eds.), *The Philosophy of Mind: Classical Problems/Contemporary Issues*. Cambridge, MA: MIT Press, 393–396.
——— ([1980] 1992b), "Discussion of Putnam's Comments," in B. Beakley and P. Ludlow (eds.), *The Philosophy of Mind: Classical Problems/Contemporary Issues*. Cambridge, MA: MIT Press, 411–422.
——— ([1988] 1992c), "From *Language and Problems of Knowledge*," in B. Beakley and P. Ludlow (eds.), *The Philosophy of Mind: Classical Problems/Contemporary Issues*. Cambridge, MA: MIT Press, 47–50.
——— (1994), "Noam Chomsky," in S. Guttenplan (ed.), *A Companion to the Philosophy of Mind*. Oxford: Blackwell, 153–167.
——— (1995), *The Minimalist Program*. Cambridge, MA: MIT Press.
——— (2000), *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.
——— (2001), "Beyond Explanatory Adequacy," *MIT Occasional Papers in Linguistics*, no. 20. MIT Department of Linguistics.
Crane, T., and D. H. Mellor (1990), "There Is No Question of Physicalism," *Mind* 99: 185–206.
Gardner, H. (1987), *The Mind's New Science: A History of the Cognitive Revolution*. New York: Basic Books.
Gould, S. J. (1980), *The Panda's Thumb: More Reflections in Natural History*. New York: W. W. Norton.
Lewontin, R. (1990), "The Evolution of Cognition," in D. N. Osherson and E. E. Smith (eds.), *An Invitation to Cognitive Science* (Vol. 3). Cambridge, MA: MIT Press, 229–246.
Newmeyer, Frederick (1986), *Linguistic Theory in America* (2nd ed.). San Diego: Academic Press.
Piatelli-Palmerini, M. (ed.) (1980), *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Cambridge, MA: Harvard University Press.
Putnam, H. (1988), *Representation and Reality*. Cambridge, MA: MIT Press.
——— ([1980] 1992), "What Is Innate and Why: Comments on the Debate," in B. Beakley and P. Ludlow (eds.), *The Philosophy of Mind: Classical Problems/Contemporary Issues*. Cambridge, MA: MIT Press.
Skinner, B. F. (1957), *Verbal Behavior*. New York: Appleton-Century-Crofts.
Smith, N. (2000), Foreword, in Chomsky, *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.

See also **Behaviorism; Cognitive Science; Innate-Acquired Distinction; Linguistics, Philosophy of; Psychology, Philosophy of; Physicalism**

CLASSICAL MECHANICS

Over the centuries classical mechanics has been a steady companion of the philosophy of science. It has played different parts, ranging (i) from positing its principles as *a priori* truths to the insight—pivotal for the formation of a modern philosophy of science—that modern physics requires a farewell to the explanatory ideal erected upon mechanics; (ii) from the physiological analyses of mechanical experiences to axiomatizations according to the strictest logical standards; and (iii) from the mechanistic philosophy to conventionalism (see Conventionalism; Determinism; Space-Time). Core structures of modern science, among them differential equations and conservation laws, as well as core themes of philosophy, among them determinism and the ontological status of theoretical terms, have emerged from this context. Two of the founders of classical mechanics, Galileo and Newton, have often been identified with the idea of modern science as a whole. Owing to its increasing conceptual and mathematical refinement during the nineteenth century, classical mechanics gave birth to the combination of formal analysis and philosophical interpretation that distinguished the modern philosophy of science from the earlier *Naturphilosophie*. The works of Helmholtz, Mach, and Poincaré molded the historical character of the scientist-philosopher that would fully bloom during the emergence of relativity theory and quantum mechanics (see Mach; Quantum Mechanics; Poincaré; Space-Time).

Mechanics became "classical" at the latest with the advent of quantum mechanics. Already relativistic field theory had challenged mechanics as the leading scientific paradigm. Consequently, present-day philosophers of science usually treat classical mechanics within historical case studies or as the first touchstone for new proposals of a general kind. Nonetheless, there are at least two lines on which classical mechanics in itself remains a worthy topic

for philosophers. On the one hand, ensuing from substantial mathematical progress during the twentieth century, classical mechanics has developed into the theory of classical dynamical systems. Among the problems of interest to formally minded philosophers of science are the relationships between the different conceptualizations, issues of stability and chaotic behavior, and whether classical mechanics is a special case of the conceptual structures characteristic of other physical theories. On the other hand, classical mechanics or semiclassical approaches continue to be applied widely by working scientists and engineers. Those real-world applications involve a variety of features that substantially differ from the highly idealized textbook models to which physicists and philosophers are typically accustomed, and they require solution strategies whose epistemological status is far from obvious.

Often classical mechanics is used synonymously with Newtonian mechanics, intending that its content is circumscribed by Newton's famous three laws. The terms **analytical mechanics** and **rational mechanics** stress the mathematical basis and theoretical side of mechanics as opposed to **practical mechanics**, which originally centered around the traditional simple machines: lever, wedge, wheel and axle, tackle block, and screw. But the domain of mechanics was being constantly enlarged by the invention of new mechanical machines and technologies. Most expositions of mechanics also include the theory of elasticity and the mechanics of continua. The traditional distinction between mechanics and physics was surprisingly long-lived. One reason was that well into the nineteenth century, no part of physics could live up to the level of formal sophistication that mechanics had achieved. Although of little importance from a theoretical point of view, it is still common to distinguish statics (mechanical systems in equilibrium) and dynamics that are subdivided into a merely geometrical part, kinematics, and kinetics.

AU:25

Ancient Greek and Early Modern Mechanics

In Greek antiquity, mechanics originally denoted the art of voluntarily causing motions against the nature of the objects moved. Other than physics, it was tractable by the methods of Euclidean geometry. Archimedes used mechanical methods to determine the center of mass of complex geometrical shapes but did not recognize them as valid geometrical proofs. His Euclidean derivation of the law of the lever, on the other hand, sparked severe criticism from Mach ([1883] 1960), who held that no mathematical derivation whatsoever could replace experience.

Pivotal for bringing about modern experimental science was Galileo's successful criticism of Aristotelian mechanics. Aristotle had divided all natural motions into celestial motions, which were circular and eternal, and terrestrial motions, which were rectilinear and finite. Each of the four elements (earth, water, fire, air) moved toward its natural place. Aristotle held that heavy bodies fell faster than light ones because velocity was determined by the relation of motive force (weight) and resistance. All forces were contact forces, such that a body moving with constant velocity was continuously acted upon by a force from the medium. Resistance guaranteed that motion remained finite in extent. Thus there was no void, nor motion in the void. Galileo's main achievement was the idealization of a constantly accelerated motion in the void that is slowed down by the resistance of the medium. This made free fall amenable to geometry. For today's reader, the geometrical derivation of the law is clumsy; the ratio of different physical quantities $v = s/t$ (where s stands for a distance, t for a time interval, and v for a velocity) was not yet meaningful. Galileo expressly put aside what caused bodies to fall and referred to experimentation. Historians and philosophers have broadly discussed what and how Galileo reasoned from empirical evidence. Interpreters wondered, in particular, whether the thought experiment establishing the absurdity of the Aristotelian position was conclusive by itself or whether Galileo had simply repackaged empirical induction in the deductive fashion of geometry.

Huygens' most important contribution to mechanics was the derivation of the laws of impact by invoking the principle of energy conservation. His solution stood at the crossroads of two traditions. On the one hand, it solved the foundational problem of Cartesian physics, the program of which was to reduce all mechanical phenomena to contact forces exchanged in collision processes. On the other hand, it gave birth to an approach based upon the concept of energy, which became an alternative to the Newtonian framework narrowly understood.

The work of Kepler has repeatedly intrigued philosophers of diverging orientations. Utilizing the mass of observational data collected by Brahe, Kepler showed that the planetary orbits were ellipses. Kepler's second law states that the line joining the sun to the planet sweeps through equal areas in equal times, and the third law states that the square of the periods of revolution of any two planets are in the same proportion as the cubes of their

semi-major axes. All three laws were merely kinematical. In his *Mysterium cosmographicum*, Kepler ([1596] 1981) identified the spacings of the then known six planets with the five platonic solids. It will be shown below that the peculiar shape of the solar system given Newton's laws of universal gravitation can be explained without reference to Kepler's metaphysical belief in numerical harmony.

On the Status of Newton's Laws

The most important personality in the history of classical mechanics was Newton. Owing to the activities of his popularizers, he became regarded as the model scientist; this admiration included a devotion to the methodology of his *Philosophiae naturalis principia mathematica* (Newton [1687, 1713] 1969), outlined in a set of rules preceding book III. But interpreters disagree whether Newton really pursued the Baconian ideal of science and licensed induction from phenomena or must be subsumed under the later descriptivist tradition that emerged with Mach ([1883] 1960) and Kirchhoff (1874). At any rate, the famous declaration not to feign hypothesis targeted the Cartesian and Leibnizian quest for a metaphysical basis of the principles of mechanics.

After the model of Euclid, the *Principia* began with eight definitions and three axioms or laws. Given Newton's empiricist methodology, interpreters have wondered about their epistemological status and the logical relations among them. Certainly, the axioms were neither self-evident truths nor mutually independent:

1. Every body continues in its state of rest or uniform motion in a right (i.e., straight) line, unless it is compelled to change that state by forces impressed upon it.
2. The change of motion is proportional to the motive force impressed and is made in the direction of the right line in which that force is impressed.
3. To every action there is always opposed an equal reaction; or the mutual actions of two bodies upon each other are always equal, and directed to contrary parts.

The proper philosophical interpretation of these three laws remained contentious until the end of the nineteenth century. Kant claimed to have deduced the first and third laws from the synthetic *a priori* categories of causality and reciprocity, respectively. The absolute distinction between rest and motion in the first law was based on absolute space and time. Within Kant's transcendental philosophy, Euclidean space-time emerged from pure intuition *a priori*. This was the stand against which twentieth-century philosophy of science rebelled, having relativity theory in its support (see Space-Time).

Admitting that the laws were suggested by previous experiences, some interpreters, including Poincaré, considered them as mere conventions (see Conventionalism). Positing absolute Euclidean space, the first law states how inertial matter moves in it. But it is impossible to obtain knowledge about space independent of everything else. Thus, the first law represents merely a criterion for choosing a suitable geometry. Mach ([1883] 1960) rejected absolute space and time as metaphysical. All observable motion was relative, and thus, at least in principle, all material objects in the universe were mutually linked. **Mach's principle**, as it became called, influenced the early development of general relativity but is still controversial among philosophers (see Barbour and Pfister 1995; Pooley and Brown 2002). To Mach's mind, the three laws were highly redundant and followed from a proper empiricist explication of Newton's definitions of mass and force. Defining force, with Newton, as an action exerted upon a body to change its state, that is, as an acceleration, the first and second laws can be straightforwardly derived. Mach rejected Newton's definition of mass as quantity of matter as a pseudo-definition and replaced it with the empirical insight that mass is the property of bodies determining acceleration. This was equivalent to the third law. Mach's criticism became an important motivation for Einstein's special theory of relativity. Yet, even the members of the Vienna Circle disagreed whether this influence was formative or Mach merely revealed the internal contradictions of the Newtonian framework (see Vienna Circle).

As to the second law, one may wonder whether the forces are inferred from the observed phenomena of motion or the motions are calculated from given specific forces. Textbooks often interpret the second law as providing a connection from forces to motion, but this is nontrivial and requires a proper superposition of the different forces and a due account of the constraints. The opposite interpretation has the advantage of not facing the notorious problem of characterizing force as an entity in its own right, which requires a distinction between fundamental forces from fictitious or inertial forces.

Book III of Newton's *Principia* introduced the law of universal gravitation, which finally unified the dynamics of the celestial and terrestrial spheres.

But, as it stood, it involved an action at a distance through the vacuum, which Newton regarded as the greatest absurdity. To Bentley he wrote: "Gravity must be caused by an Agent acting according to certain Laws; but whether this Agent be material or immaterial, I have left to the Consideration of my readers" (Cohen 1958, 303). Given the law of universal gravity, the peculiar shape of the solar system was a matter of initial conditions, a fact that Newton ascribed to the contrivance of a voluntary agent.

The invention of the calculus was no less important for the development of mechanics than was the law of gravitation. But its use in the *Principia* was not consistent and intertwined with geometrical arguments, and it quickly turned out that Leibniz's version of the calculus was far more elegant. That British scientists remained loyal to Newton's fluxions until the days of Hamilton proved to be a substantial impediment to the use of calculus in science.

Celestial Mechanics and the Apparent Triumph of Determinism

No other field of mechanics witnessed greater triumphs of prediction than celestial mechanics: the return of Halley's comet in 1758; the oblateness of the Earth in the 1740s; and finally the discovery of Neptune in 1846 (cf. Grosser 1962). However, while Neptune was found at the location that the anomalies in the motion of Uranus had suggested, Le Verrier's prediction of a planet Vulcan to account for the anomalous perihelion motion of Mercury failed. Only general relativity would provide a satisfactory explanation of this anomaly (see Space-Time).

But in actual fact little follows from Newton's axioms and the inverse square law of gravitation alone. Only the two-body problem can be solved analytically by reducing it to a one-body problem for relative distance. The three-body problem requires approximation techniques, even if the mass of one body can be neglected. To ensure the convergence of the respective perturbation series became a major mathematical task. In his lunar theory, Clairaut could derive most of the motion of the lunar apsides from the inverse-square law, provided the approximation was carried far enough. But d'Alembert warned that further iterations might fail to converge; hence subsequent analysts calculated to higher and higher orders. D'Alembert's derivation of the precession and nutation of the Earth completed a series of breakthroughs around 1750 that won Newton's law a wide acceptance. In 1785, Laplace explained the remaining chief anomalies in the solar system and provided a (flawed) indirect proof for the stability of the solar system from the conservation of angular momentum (cf. Wilson 1995).

One might draw an inductivist lesson from this history and conceive the increased precision in the core parameters of the solar system, above all the planetary masses, as a measure of explanatory success for the theory containing them (cf. Harper and Smith 1995). Yet, the historically more influential lesson for philosophers consisted in the ideal of **Laplace's Demon**, the intellect that became the executive officer of strict determinism:

> Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it—an intelligence sufficiently vast to submit these data to analysis—it would embrace in the same formula the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain, and the future, as the past, would be present to its eyes.
>
> (Laplace [1795] 1952, 4)

But the Demon is threatened by idleness in many ways. If the force laws are too complex, determinism becomes tautologous; if Newton's equations cannot be integrated, perturbative and statistical strategies are mandatory; calculations could still be too complex; exact knowledge of the initial state of a system presupposes that the precision of measurement could be increased at will.

A further idealization necessary to make the Laplacian ideal thrive was the **point mass approach** of Boscovich ([1763] 1966). But some problems cannot be solved in this way—for instance, whether a point particle moving in a head-on orbit directed at the origin of a central force is reflected by the singularity or goes right through it. In many cases, celestial mechanics treats planets as extended bodies or gyroscopes rather than point masses. Nonrigid bodies or motion in resisting media require even further departures from the Laplacian ideal. As Mark Wilson put it, "applied mathematicians are often forced to pursue roundabout and shakily rationalized expedients if any progress is to be made" (2000, 296).

Only some of these expedients can be mathematically justified. This poses problems for the relationship of mathematical and physical ontology. Often unsolvable equations are divided into tractable satellite equations. Continuum mechanics is a case in point. The Navier-Stokes equations are virtually intractable by analytic methods; Prandtl's boundary layer theory splits a flow in a pipe into

one equation for the fluid's boundary and one for the middle of the flow. Prandtl's theory failed in the case of turbulence, while statistical investigations brought useful results. Accordingly, in this case predictability required abandoning determinism altogether years before the advent of quantum mechanics (cf. von Mises 1922).

From Conserved Quantities to Invariances: Force and Energy

The framework of Newton's laws was not the only conception of mechanics used by 1800 (cf. Grattan-Guinness 1990). There were purely algebraic versions of the calculus based on variational problems developed by Euler and perfected in Lagrange's *Analytical Mechanics* ([1788] 1997). Engineers developed a kind of energy mechanics that emerged from Coulomb's friction studies. Lazare Carnot developed it into an alternative to Lagrange's reduction of dynamics to equilibrium, that is, to statics. Dynamics should come first, and engineers had to deal with many forces that did not admit a potential function.

The nature of energy—primary entity or just inferred quantity—was no less in dispute than that of force. Both were intermingled in content and terminology. The eighteenth century was strongly influenced by the *vis viva* controversy launched by Leibniz. What was the proper quantity in the description of mechanical processes: mv or mv^2 (where m = mass and v = velocity)? After the Leibniz-Clarke correspondence, the issue became associated with atomism and the metaphysical character of conserved quantities. In the nineteenth century, 'energy,' or work, became a universal principle after the discovery of the mechanical equivalent of heat. Helmholtz gave the principle a more general form based upon the mechanical conception of nature. For many scientists this suggested a reduction of the other domains of physical science to mechanics. But this program failed in electrodynamics.

Through the works of Ostwald and Helm, the concept of energy became the center of a movement that spellbound German-speaking academia from the 1880s until the 1900s. But, as Boltzmann untiringly stressed, energeticists obtained the equations of motion only by assuming energy conservation in each spatial direction. But why should Cartesian coordinates have a special meaning?

Two historico-critical studies of the energy concept set the stage for the influential controversy between Planck and Mach (1908–1910). While Mach ([1872] 1911) considered the conservation of work as an empirical specification of the instinctive experience that perpetual motion is impossible, Planck (1887) stressed the independence and universality of the principle of energy conservation. Planck later lauded Mach's positivism as a useful antidote against exaggerated mechanical reductionism but called for a stable world view erected upon unifying variational principles and invariant quantities.

Group theory permitted mathematical physicists to ultimately drop any substantivalist connotation of conserved quantities, such as energy. The main achievement was a theorem of Emmy Noether that identified conserved quantities, or constants of motion, with invariances under one-parameter group transformations (cf. Arnold 1989, 88–90).

Variational Principles and Hamiltonian Mechanics

In 1696–7 Johann Bernoulli posed the problem of finding the curve of quickest descent between two points in a homogeneous gravitational field. While his own solution used an analogy between geometrical optics and mechanics, the solutions of Leibniz and Jacob Bernoulli assumed that the property of minimality was present in the small as in the large, arriving thus at a differential equation. The title of Euler's classic treatise describes the scope of the variational calculus: *The Method of Finding Curved Lines That Show Some Property of Maximum or Minimum* (1744). The issue of minimality sparked philosophical confusion. In 1746 Pierre Moreau de Maupertuis announced that in all natural processes the quantity $\int vds = \int v^2 dt$ attains its minimum; he interpreted this as a formal teleological principle that avoided the outgrowths of the earlier physicotheology of Boyle and Bentley. Due to his defective examples and a priority struggle, the principle lost much of its credit. Lagrange conceived of it simply as an effective problem-solving machinery. There was considerable mathematical progress during the nineteenth century, most notably by Hamilton, Gauss, and Jacobi. Only Weierstrass found the first sufficient condition for the variational integral to actually attain its minimum value. In 1900, Hilbert urged mathematicians to systematically develop the variational calculus (see Hilbert, David). He made action principles a core element of his axiomatizations of physics. Hilbert and Planck cherished the principles' applicability, after appropriate specification, to all fields of reversible physics; thus they promised a formal unification instead of the discredited mechanical reductionism.

There exist differential principles, among them d'Alembert's principle and Lagrange's principle of virtual work, that reduce dynamics to statics, and integral action principles that characterize the actual dynamical evolution over a finite time by the stationarity of an integral as compared with other possible evolutions. Taking M_q as the space of all possible motions $q(t)$ between two points in configuration space, the action principle states that the actual motion q extremizes the value of the integral $W[q] = \int_a^b F(t, q(t), \dot{q}(t))dt$ in comparison with all possible motions, the varied curves $(q + \delta q) \in M_q$ (the endpoints of the interval $[a, b]$ remain fixed).

If one views the philosophical core of action principles in a temporal teleology, insofar as a particle's motion between a and b is also determined by the fixed state, this contradicts the fact that almost all motions that can be treated by way of action principles are reversible. Yet, already the mathematical conditions for an action principle to be well defined suggest that the gist of the matter lies in its global and modal features. Among the necessary conditions is the absence of conjugate points between a and b through which all varied curves have to pass. Sufficient conditions typically involve a specific field embedding of the varied curves; also the continuity properties of the $q \in M_q$ play a role. Thus initial and final conditions have to be understood in the same vein as boundary conditions for partial differential equations AU:26 that have to be specified beforehand so that the solution cannot be grown stepwise from initial data.

There are also different ways of associating the possible motions. In Hamilton's principle, F is the Lagrangian $L = T - V$, where T is the kinetic and V the potential energy; time is not varied such that all possible motions take equal time but correspond to higher energies than the actual motion. For the original principle of least action, F equals T, and one obtains the equations of motion only by assuming energy conservation, such that at equal total energy the varied motions take a longer time than the actual ones.

Butterfield (2004) spots three types of modality in the sense of David Lewis (1986) here. While all action principles involve a modality of changed initial conditions and changed problems, Hamilton's principle also involves changed laws because the varied curves violate energy conservation. Energeticists and Mach ([1883] 1960) held that u is uniquely determined within M_u as compared with the other motions that appear pairwise or with higher degeneracy, but they disagreed whether one could draw conclusions about modal characteristics. Albeit well versed in their mathematical intricacies, logical empiricists treated action principles with neglect in order to prevent an intrusion of metaphysics (Stöltzner 2003).

The second-order Euler-Lagrange equations, which typically correspond to the equations established by way of Newton's laws, can be reformulated as a pair of first-order equations, Hamilton's equations, or a single partial differential equation, *viz.*, the **Hamilton-Jacobi equation**. Hamilton's equations emerge from the so-called Legendre transformation, which maps configuration space $(q, \dot{q})$ into phase space (q, p), thus transforming the derivative of position $\dot{q}$ into momentum p. If AU:27 this transformation fails, constraints may be present. The core property of Hamilton's equations is their invariance under the so-called **canonical transformations**. The canonical transformation $(q, p) \rightarrow (Q, P)$ that renders the Hamiltonian $H(Q,P) = T + V$ equal to zero leads to the Hamilton-Jacobi equation. Its generator $W = S - E_t$ (where E_t is the total energy) can be interpreted as an action AU:28 functional corresponding to moving wave fronts in ordinary space that are orthogonal to the extremals of the variational problem (cf. Lanczos 1986).

The analogy between mechanics and geometric optics is generic and played an important role in Schrödinger's justification of his wave equation. Hamilton-Jacobi theory was also a motivation for Bohm's reformulation of the de Broglie pilot wave theory (see Quantum Mechanics). For periodical motions one can use S to generate a canonical transformation that arrives at action and angle variables (J, ω), where $J = \oint pdq$. This integral was quantized in the older Bohr-Sommerfeld quantum theory. The space of all possible motions associated with a variational problem M_q can be considered as an ensemble of trajectories. Arguing that in quantum theory all possible motions are realized with a certain possibility provides some intuitive motivation for the Feynman path integral approach (see Quantum Field Theory).

Variational principles played a central role in several philosophically influential treatises of mechanics in the late nineteenth century. Most radical was that of Heinrich Hertz ([1894] 1956), who used Gauss's (differential) principle of least constraints to dispense with the notion of force altogether. There were no single mass points but only a system of mass points connected by constraints. Hertz's problem was to obtain a geometry of straight line for this system of mass points. This task was complicated by Hertz's Kantian preference for Euclidean

geometry as a basis for the non-Euclidean geometry of mass points. Contemporaries deemed Hertz's geometrization of mechanics a "God's eye view." Boltzmann (1897 and 1904) praised its coherence but judged Hertz's picture as inapplicable to problems easily tractable by means of forces.

AU:29

Hertz held that theoretical pictures had to be logically permissible, empirically correct, and appropriate (that is, sufficiently complete and simple). Boltzmann criticized permissibility as an unwarranted reliance upon *a priori* laws of thought and held instead that pictures were historically acquired and corroborated by success. Around 1900, treating theories (e.g., atomism) as pictures represented an alternative to mechanical reductionism and positivist descriptivism, until the idea of coordinative definitions between symbolic theory and empirical observations won favor (see Vienna Circle).

Mathematical Mechanics and Classical Dynamical Systems

After classical mechanics was finally dethroned as the governing paradigm of physics, its course became largely mathematical in kind and was strongly influenced by concepts and techniques developed by the new-frontier theories of physics: **relativity theory** and **quantum physics**. There was substantial progress in the variational calculus, but the main inspirations came from differential geometry, group theory, and topology, on the one hand, and probability theory and measure theory, on the other. This has led to some rigorous results on the n-body problem. The advent of the modern computer not only opened a new era of celestial mechanics, it also revealed that chaotic behavior, ignored by physicists in spite of its early discovery by Poincaré, occurred in a variety of simple mechanical systems.

Geometrization has drastically changed the appearance of classical mechanics. Configuration space and phase space have become the tangent and cotangent bundles on which tensors, vector fields, and differential forms are defined. Coordinates have turned into charts, and the theory of differentiable manifolds studies the relationships between the local level, where everything looks Euclidean, and the global level, where topological obstructions may arise. The dynamics acting on these bundles are expressed in terms of flows defined by vector fields, which gives a precise meaning to variations. The picture of flows continues Hamilton-Jacobi theory. This elucidates that the intricacies of variational calculus do not evaporate; they transform into obstructions of and conditions for the application of the whole geometrical machinery, e.g., how far a local flow can be extended. The constants of the motion define invariant submanifolds that restrict the flow to a manifold of lower dimension; they act like constraints. If a Hamiltonian system has, apart from total energy, enough linearly independent constants of motion and if their Poisson brackets $\{F, G\} = \sum_{i=1}^{m} (\partial F/\partial q_i \partial G/\partial p_i - \partial G/\partial q_i \partial F/\partial p_i)$ mutually vanish, the system is integrable (the equations of motion can be solved) and the invariant submanifold can be identified with a higher-dimensional torus.

The geometrical structure of Hamiltonian mechanics is kept together by the symplectic form $\omega = dq^i \wedge dp_i$ that is left invariant by canonical transformations. It plays a role analogous to the metric in relativity theory. The skew-symmetric product $\wedge$ of forms and the exterior derivative d are the main tools of the Cartan calculus and permit an entirely coordinate-free formulation of dynamics. Many equations thus drastically simplify, but quantitative results still require the choice of a specific chart.

Mathematical progress went hand in hand with a shift of emphasis from quantitative to qualitative results that started with Poincaré's work on the n-body problem. Some deep theorems of Hamiltonian mechanics become trivialities in this new language, among them the invariance of phase space volume (Liouville's theorem) and the fact that almost all points in a phase space volume eventually—in fact, infinitely often—return arbitrarily close to their original position (Poincaré recurrence) (see also Statistical Mechanics).

For the small number of mass points characteristic of celestial mechanics, there has been major progress along the lines of the questions asked by Thirring (1992, 6): "Which configurations are stable? Will collisions ever occur? Will particles ever escape to infinity? Will the trajectory always remain in a bounded region of phase space?" In the two-body problem, periodic elliptic motions, head-on collisions, and the hyperbolic and parabolic trajectories leading to infinity are neatly separated by initial data.

For the three-body problem the situation is already complex; there do not exist sufficient constants of motion. Two types of exact solutions were quickly found: The particles remain collinear (Euler) or they remain on the vertices of an equilateral triangle (Lagrange). The equilateral configuration is realized by the Trojan asteroids, Jupiter and

the sun. In general, however, even for a negative total energy, two bodies can come close enough to expel the third to infinity. In the four-body problem, there exist even collinear trajectories on which a particle might reach spatial infinity in a finite time (Mather and McGehee 1975). Five bodies, one of them lighter than the others, can even be arranged in such a manner that this happens without any previous collisions because the fifth particle oscillates faster and faster between the two escaping planes, in each of which two particles rotate around one another (Xia 1992; Saari 2005). Both examples blatantly violate special relativity. But rather than hastily resort to arguments of what is 'physical,' the philosopher should follow the mathematical physicist in analyzing the logical structure of classical mechanics, including collisions and other singularities.

Can an integrable system remain stable under perturbation? In the 1960s Kolmogorov, Arnold, and Moser (KAM) gave a very general answer that avails itself of the identification of integrable systems and invariant tori. The **KAM theorem** shows that sufficiently small perturbations deform only the invariant tori. If the perturbation increases, the tori in resonance with it break first; or more precisely, the more irrational the ratio of the frequency of an invariant torus (in action and angle variables) and the frequency of the perturbation, the more stable is the respective torus. If all invariant tori are broken, the system becomes chaotic (cf. Arnold 1989; Thirring 1992). The KAM theorem also provides the rigorous basis of Kepler's association of planetary orbits and platonic solids. The ratios of the radii of platonic solids to the radii of inscribed platonic solids are irrational numbers of a kind that is badly approximated by rational numbers. And, indeed, among the asteroids between Mars and Jupiter, one finds significant gaps for small ratios of an asteroid's revolution time to that of the perturbing Jupiter.

If all invariant tori have broken up, the only remaining constant of motion is energy, such that the system begins to densely cover the energy shell and becomes ergodic. No wonder that concepts of statistical physics, among them ergodicity, entropy, and mixing, have been used to classify chaotic behavior, even though chaotic behavior does not succumb to statistical laws. They are supplemented by topological concepts, such as the Hausdorff dimension, and concepts of dynamical systems theory, such as bifurcations (nonuniqueness of the time evolution) and attractors (in the case of dissipative systems where energy is no longer conserved).

MICHAEL STÖLTZNER

References

Arnold, Vladimir I (1989), *Mathematical Methods of Classical Mechanics*, 2nd ed. New York and Berlin: Springer.

Barbour, Julian B., and Herbert Pfister (eds.) (1995), *Mach's Principle: From Newton's Bucket to Quantum Gravity*. Boston and Basel: Birkhäuser.

Boltzmann, Ludwig (1897 and 1904), *Vorlesungen über die Principe der Mechanik*. 2 vols. Leipzig: Barth.

Boscovich, Ruder J. ([1763] 1966), *A Theory of Natural Philosophy*. Cambridge, MA: MIT Press.

Butterfield, Jeremy (2004), "Some Aspects of Modality in Analytical Mechanics," in Michael Stöltzner and Paul Weingartner (eds.), *Formale Teleologie und Kausalität*. Paderborn, Germany: Mentis.

Cohen, I. Bernhard (ed.) (1958), *Isaac Newton's Papers and Letters on Natural Philosophy*. Cambridge, MA: Harvard University Press.

Euler, Leonhard (1744), *Methodus Inveniendi Lineas Curvas Maximi Minimive Proprietate Gaudentes sive Solutio Problematis Isoperimetrici Latissimo Sensu Accepti*. Lausanne: Bousquet.

Grattan-Guinness, Ivor (1990), "The Varieties of Mechanics by 1800," *Historia Mathematica* 17: 313–338.

Grosser, Morton (1962), *The Discovery of Neptune*. Cambridge, MA: Harvard University Press.

Harper, William, and George E. Smith (1995), "Newton's New Way of Inquiry," in J. Leplin (ed.), *The Creation of Ideas in Physics: Studies for a Methodology of Theory Construction*. Dordrecht, Netherlands: Kluwer, 113–166.

Hertz, Heinrich ([1894] 1956), *The Principles of Mechanics: Presented in a New Form*. New York: Dover Publications.

Kepler, Johannes ([1596] 1981), *Mysterium cosmographicum*. New York: Abaris.

Kirchhoff, Gustav Robert (1874), *Vorlesungen über analytische Mechanik*. Leipzig: Teubner.

Lagrange, Joseph Louis ([1788] 1997), *Analytical Mechanics*. Dordrecht, Netherlands: Kluwer.

Lanczos, Cornelius (1986), *The Variational Principles of Mechanics*. New York: Dover.

Laplace, Pierre Simon de ([1795] 1952), *A Philosophical Essay on Probabilities*. New York: Dover.

Lewis, David (1986), *On the Plurality of Worlds*. Oxford: Blackwell.

Mach, Ernst ([1872] 1911), *History and Root of the Principle of the Conservation of Energy*. Chicago: Open Court.

——— ([1883] 1960), *The Science of Mechanics: A Critical and Historical Account of Its Development*. La Salle, IL: Open Court. German first edition published by Brockhaus, Leipzig, 1883.

Mather, John, and Richard McGehee, "Solutions of the Collinear Four-Body Problem Which Become Unbounded in Finite Time," in Jürgen Moser (ed.), *Dynamical Systems Theory and Applications*. New York: Springer, 573–587.

Newton, Isaac ([1687, 1713] 1969), *Mathematical Principles of Natural Philosophy*. Translated by Andrew Motte (1729) and revised by Florian Cajori. 2 vols. New York: Greenwood.

Planck, Max (1887), *Das Princip der Erhaltung der Energie*, 2nd ed. Leipzig: B. G. Teubner.

Pooley, Oliver, and Harvey Brown (2002), "Relationalism Rehabilitated? I: Classical Mechanics," *British Journal for the Philosophy of Science* 53: 183–204.

AU:30

Saari, Donald G. (2005), *Collisions, Rings, and Other Newtonian* N-*Body Problems*. Providence, RI: American Mathematical Society.

Stöltzner, Michael (2003), "The Principle of Least Action as the Logical Empiricists' Shibboleth," *Studies in History and Philosophy of Modern Physics* 34: 285–318.

Thirring, Walter (1992), *A Course in Mathematical Physics 1: Classical Dynamical Systems*, 2nd ed. Translated by Evans M. Harrell. New York and Vienna: Springer.

von Mises, Richard (1922), "Über die gegenwärtige Krise der Mechanik," *Die Naturwissenschaften* 10: 25–29.

Wilson, Curtis (1995), "The Dynamics of the Solar System," in Ivor Grattan-Guinness (ed.), *Companion Encyclopedia of the History and Philosophy of the Mathematical Sciences*. London: Routledge.

Wilson, Mark (2000), "The Unreasonable Uncooperativeness of Mathematics in the Natural Sciences," *The Monist* 83 (2000): 296–314. See also his very instructive entry "Classical Mechanics" in the *Routledge Encyclopedia of Philosophy*.

Xia, Zhihong (1992), "The Existence of Noncollision Singularities in Newtonian Systems," *Annals of Mathematics* 135: 411–468.

COGNITIVE SCIENCE

Cognitive science is a multidisciplinary approach to the study of cognition and intelligence that emerged in the late 1950s and 1960s. "Core" cognitive science holds that the mind/brain is a kind of computer that processes information in the form of mental representations. The major disciplinary participants in the cognitive science enterprise are psychology, linguistics, neuroscience, computer science, and philosophy. Other fields that are sometimes included are anthropology, education, mathematics, biology, and sociology.

There has been considerable philosophical discussion in recent years that relates, in one way or another, to cognitive science. Arguably, not all of this discussion falls within the tradition of philosophy of science. There are two fundamentally different kinds of question that philosophers of science typically raise about a specific scientific field: "external" questions and "internal" questions. If one assumes that a scientific field provides a framework of inquiry, external questions will be about that framework as a whole, from some external point of view, or about its relation to other scientific frameworks. In contrast, an internal question will be one that is asked *within* the framework, either with respect to some entity or process that constitutes part of the framework's foundations or with respect to some specific theoretical/empirical issue that scientists committed to that framework are addressing. Philosophers have dealt extensively with both external and internal questions associated with cognitive science.

Key External Questions

There are three basic groups of external questions: those concerning the nature of a scientific field X, those concerning the relations of X to other scientific fields, and those concerning the scientific merits of X. An interesting feature of such discussions is that they often draw on, and sometimes even contribute to, the literature on a relevant prior meta-question. For example, discussions concerning the scientific nature of a particular X (e.g., cognitive science) may require prior consideration of the question, What is the best way to characterize X (in general) scientifically (e.g., in terms of its theories, explanations, paradigm, research program)?

What Is Cognitive Science?

There are many *informal* descriptions of cognitive science in the literature, not based on any serious consideration of the relevant meta-question (e.g., Simon 1981; Gardner 1985). The only systematic formal treatment of cognitive science is that of von Eckardt (1993), although there have been attempts to describe aspects of *cognitive psychology* in terms of Kuhn's notion of a paradigm. Rejecting Kuhn's notion of a paradigm as unsuitable for the characterization of an *immature* field, von Eckardt (1993) proposes, as an alternative, the notion of a *research framework*. A research framework consists of four sets of elements D, Q, SA, MA, where D is a set of assumptions that provide a pretheoretical

specification of the domain under study; *Q* is a set of basic empirical research questions, formulated pretheoretically; *SA* is a set of substantive assumptions that embody the approach being taken in answering the basic questions and that constrain possible answers to those questions; and *MA* is a set of methodological assumptions.

One can think of what Kuhn (1970) calls "normal science" as problem-solving activity. The fundamental problem facing the community of researchers committed to a research framework is to answer each of the basic empirical questions of that framework, subject to the following constraints:

1. Each answer is scientifically acceptable in light of the scientific standards set down by the shared and specific methodological assumptions of the research framework.
2. Each answer is consistent with the substantive assumptions of the research framework.
3. Each answer to a theoretical question makes significant reference to the entities and processes posited by the substantive assumptions of the research framework.

According to von Eckardt (1993), cognitive science consists of a *set* of overlapping research frameworks, each concerned with one or another aspect of human cognitive capacities. Arguably, the *central* research framework concerns the study of adult, normal, typical cognition, with subsidiary frameworks focused on individual differences, group differences (e.g., expert vs. novice, male vs. female), cultural variation, development, pathology, and neural realization. The description offered in von Eckardt (1993), summarized in Table 1, is intended to be of only the central research framework. In addition, it is claimed to be a rational reconstruction as well as transdisciplinary, that is, common to cognitive scientists of all disciplines. Von Eckardt claims that research frameworks can evolve and that the description in question reflects commitments of the cognitive science community at only one period of its history (specifically, the late 1980s/early 1990s).

Cognitive science is a very complex and rapidly changing field. In light of this complexity and change, von Eckardt's original reconstruction should, perhaps, be modified as follows. First, it needs to be acknowledged that there exists a fundamental disagreement within the field as to its domain. The narrow conception, embraced by most psychologists and linguists, is that the domain is *human cognition;* the broad conception, embraced by artificial intelligence researchers, is that the domain is *intelligence in general*. (Philosophers seem to be about evenly split.) A further area of disagreement concerns whether cognitive science encompasses only cognition/intelligence or includes all aspects of mind, including touch, taste, smell, emotion, mood, motivation, personality, and motor skills. One point on which there now seems to be unanimity is that cognitive science must address the phenomenon of consciousness.

Second, a modified characterization of cognitive science must describe it as evolving not only with respect to the **computational assumption** (from an exclusive focus on symbol systems to the inclusion of connectionist devices) but also with respect to the role of neuroscience. Because cognitive science originally emerged from cognitive psychology, artificial intelligence research, and generative linguistics, in the early years neuroscience was often relegated to a secondary role. Currently, most nonneural cognitive scientists believe that research on the mind/brain should proceed in an interactive way—simultaneously both top-down and bottom-up. Ironically, a dominant emerging view of cognitive neuroscience seems to be that it, rather than cognitive *science*, will be the locus of an interdisciplinary effort to develop, from the bottom up, a computational or information-processing theory of the mind/brain.

There are also research programs currently at the periphery of cognitive science—what might be called alternative cognitive science. These are research programs that investigate some aspect of cognition or intelligence but whose proponents reject one or more of the major guiding or methodological assumptions of mainstream cognitive science. Three such programs are research on situated cognition, artificial life, and the dynamical approach to cognition.

Interfield Relations Within Cognitive Science

The second group of external questions typically asked by philosophers of some particular scientific field concerns the relation of that field to other scientific fields. Because cognitive science is itself multidisciplinary, the most pressing interfield questions arise about the relationship of the various subdisciplines *within* cognitive science. Of the ten possible two-place relations among the five core disciplines of cognitive science, two have received the most attention from philosophers: relations between linguistics and psychology (particularly psycholinguistics) and relations between cognitive psychology and neuroscience.

Discussion of the relation between linguistics and psychology has addressed primarily Chomsky's

Table 1. The central research framework of cognitive science

Domain-specifying assumptions

D1 (Identification assumption): The domain consists of the human cognitive capacities.
D2 (Property assumption): Pretheoretically conceived, the human cognitive capacities have a number of *basic general properties:*

- Each capacity is *intentional*; that is, it involves states that have content or are "about" something.
- Virtually all of the capacities are *pragmatically evaluable;* that is, they can be exercised with varying degrees of success.
- When successfully exercised, each of the evaluable capacities has a certain *coherence* or cogency.
- Most of the evaluable capacities are *reliable*; that is, typically, they are exercised successfully (at least to some degree) rather than unsuccessfully.
- Most of the capacities are *productive*; that is, once a person has the capacity in question, he or she is typically in a position to manifest it in a practically unlimited number of novel ways
- Each capacity involves one or more *conscious* states.[a]

D3 (Grouping assumption): The cognitive capacities of the normal, typical adult make up a theoretically coherent set of phenomena, or a *system*. This means that with sufficient research, it will be possible to arrive at a set of answers to the basic questions of the research framework that constitute a unified theory and are empirically and conceptually acceptable.

Basic questions

Q1: For the normal, typical adult, what precisely is the human capacity to _____?
Q2: In virtue of what does a normal, typical adult have the capacity to _____ (such that this capacity is intentional, pragmatically evaluable, coherent, reliable, and productive and involves consciousness?[a])
Q3: How does a normal, typical adult exercise his or her capacity to _____?
Q4: How does the capacity to _____ of the normal, typical adult interact with the rest of his or her cognitive capacities?

Substantive assumptions

SA1: The computational assumption

C1: (Linking assumption): The human, cognitive mind/brain is a computational device (computer); hence, the human cognitive capacities consist, to a large extent, of a system of computational capacities.
C2: (System assumption): A computer is a device capable of automatically inputting, storing, manipulating, and outputting information in virtue of inputting, storing, manipulating, and outputting representations of that information. These information processes occur in accordance with a finite set of rules that are effective and that are, in some sense, in the machine itself.

SA2: The representational assumption

R1 (Linking assumption): The human, cognitive mind/brain is a representational device; hence, the human cognitive capacities consist of a system of representational capacities.
R2 (System assumption): A representational device is a device that has states or that contains within it entities that are representations. Any representation will have four aspects essential to its being a representation: (1) It will be realized by a representation bearer, (2) it will represent one or more representational objects, (3) its representation relations will be grounded somehow, and (4) it will be interpretable by (will function as a representation *for*) some currently existing interpreter. In the mind/brain representational system, the nature of these four aspects is constrained as follows:

- R2.1 (The representation bearer): The representation bearer of a mental representation is a computational structure or process, considered purely formally.
  (a) This structure or process has *constituent structure.*[b]
- R2.2 (The representational content): Mental representations have a number of semantic properties.[c]
  (a) They are semantically *selective.*
  (b) They are semantically *diverse.*
  (c) They are semantically *complex.*
  (d) They are semantically *evaluable.*
  (e) They have *compositional* semantics.[b]

(*Continued*)

Table 1. (*Continued*)

- R2.3 (The ground): The ground of a mental representation is a property or relation that determines the fact that the representation has the object (or content) it has.
  a) This ground is *naturalistic* (that is, nonsemantic and nonintentional).
  b) This ground may consist of either internal or external factors. However, any such factor must satisfy the following restriction: If two token representations have different grounds, and this ground difference determines a difference in content, then they must also differ in theircausal powers to produce relevant effects across nomologically possible contexts.[b]
- R2.4 (The interpretant): Mental representations are significant for the person in whose mind they "reside." The interpretant of a mental representation *R* for some subject *S* consists of the set of all possible determinate computational processes contingent upon entertaining *R* in *S*.

Methodological assumptions

M1: Human cognition can be successfully studied by focusing exclusively on the individual cognizer and his or her place in the natural environment. The influence of society or culture on individual cognition can always be explained by appealing to the fact that this influence is mediated through individual perception and representation.
M2: The human cognitive capacities are sufficiently autonomous from other aspects of mind such as affect and personality that, to a large extent, they can be successfully studied in isolation.
M3: There exists a partitioning of cognition in general into individual cognitive capacities such that each of these individual capacities can, to a large extent, be successfully studied in isolation from each of the others.
M4: Although there is considerable variation in how adult human beings exercise their cognitive capacities, it is meaningful to distinguish, at least roughly, between "normal" and "abnormal" cognition.
M5: Although there is considerable variation in how adult human beings exercise their cognitive capacities, adults are sufficiently alike when they cognize that it is meaningful to talk about a "typical" adult cognizer and it is possible to arrive at generalizations about cognition that hold (at least approximately) for all normal adults.
M6: The explanatory strategy of cognitive science is sound. In particular, answers to the original basic questions can, to a large extent, be obtained by answering their narrow information-processing counterparts (that is, those involving processes purely "in the head").
M7: In choosing among alternative hypothesized answers to the basic questions of the research framework, one should invoke the usual canons of scientific methodolology. That is, ultimately, answers to the basic questions should be justified on empirical grounds.
M8: A complete theory of human cognition will not be possible without a substantial contribution from each of the subdisciplines of cognitive science.
M9: Information-processing answers to the basic questions of cognitive science are constrained by the findings of neuroscience.[c]
M10: The optimal research strategy for developing an adequate theory of the cognitive mind/brain is to adopt a coevolutionary approach—that is, to develop information-processing answers to the basic questions of cognitive science on the basis of empirical findings from both the nonneural cognitive sciences and the neurosciences.[c]
M11: Information-processing theories of the cognitive mind/brain can explain certain features of cognition that cannot be explained by means of lower-level neuroscientific accounts. Such theories are thus, in principle, explanatorily ineliminable.[c]

Key: [a]Not in von Eckardt (1993); [b]Controversial; [c]Included on normative grounds (given the commitment of cognitive science to *X*, they should be committed to *Y*)

AU:45

claim that the aims of linguistics are, first, to develop hypotheses (in the form of generative grammars) about what the native speaker of a language *knows* (tacitly) or that capture the native speaker's linguistic *competence* and, second, to develop a theory of the innate constraints on language ("universal grammar") that the child brings to bear in learning any given language. Philosophers have focused on the relation of competence models to so-called performance models, that is, models developed by psychologists to describe the information processes involved in understanding and producing language. Earlier discussion focused on syntax; more recent debates have looked at semantics.

There are several views. One, favored by Chomsky himself, is that a representation of the grammar of a language *L* constitutes a *part* of the mental apparatus involved in the psychological processes underlying language performance and, hence, is causally implicated in the production of that performance (Chomsky and Katz 1974). A second, suggested by Marr (1982) and others, is that

competence theories describe a native speaker's linguistic input-output *capacity* (Marr's "computational" level) without describing the processes underlying that capacity (Marr's "algorithmic" level). For example, the capacity to understand a sentence *S* can be viewed as the capacity that maps a phonological representation of *S* onto a semantic representation *S*. Both views have been criticized. Against the first, it has been argued that because of the linguist's concern with simplicity and generality, it is unlikely that the formal structures utilized by optimal linguistic theories will be isomorphic to the internal representations posited by psycholinguists (Soames 1984). A complementary point is that because processing models are sensitive to architectural and resource constraints, the ways in which they implement syntactic knowledge have turned out to be much less transparent than followers of Chomsky had hoped (Mathews 1991). At best, they permit a psycholinguistic explanation of why a particular set of syntactic rules and principles correctly characterize linguistic competence. Against the second capacity view, it has been argued that grammars constitute idealizations (for example, there are no limitations on the length of permissible sentences or on their degree of internal complexity) and so do not describe actual speakers' linguistic capacities (Franks 1995; see Philosophy of Psychology; Neurobiology).

Internal questions have been raised about both foundational assumptions and concepts and particular theories and findings within cognitive science. Foundational discussions have focused on the core substantive assumptions: (1) The cognitive mind/brain is a computational system and (2) it is a representational system.

The Computational Assumption

Cognitive science assumes that the mind/brain is a kind of computer. But what *is* a computer? To date, two *kinds* of computer have been important to cognitive science modeling—*classical* machines ("conventional," "von Neumann," or "symbol system"), and *connectionist* machines ("parallel distributed processing"). Classical machines have an architecture similar to ordinary personal computers (PCs). There are separate components for processing and memory, and processing occurs by the manipulation of data structures. In contrast, connectionist computers are more like brains, consisting of interconnected networks of neuron-like elements. Philosophers interested in the computational assumption have focused on two questions: What is a computer *in general* (such that both classical and connectionist machines count as computers), and is there any reason to believe, at the current stage of research, that human mind/brains are one sort of computer rather than another? The view that the human mind/brain is, or is importantly like, a classical computer is called classicism; the view that it is, or is importantly like, a connectionist computer is called connectionism.

The theory of computation in mathematics defines a number of different kinds of *abstract* machines in terms of the sets of functions they can compute. Of these, the most relevant to cognitive science is the universal Turing machine, which, according to the Church-Turing thesis, can compute any function that can be computed by an *effective* method, that is, a method specified by a finite number of exact instructions, in a finite number of steps, carried out by a human unaided by any machinery except paper and pencil and demanding no insight or ingenuity. Many philosophers and cognitive scientists think that the notion of a computer relevant to cognitive science can be adequately captured by the notion of a Turing machine. In fact, that is not the case. To say that a computer (and hence the mind/brain) is simply a device *equivalent* to a universal Turing machine says nothing about the device's internal structure, and to say that a computer (and hence the mind/brain) is an *implementation* of a Turing machine seems flatout false. There are many dissimilarities. A Turing machine is in only one state at a time, while humans are, typically, in many mental or brain states simultaneously. Further, the memory of a Turing machine is a "tape" with only a simple linear structure, while human memory appears to have a complex multidimensional structure.

An alternative approach is to provide an architectural characterization, but at a fairly high level of abstraction. For example, von Eckardt (1993, 114) claims that cognitive science's computational assumption takes a computer to be "a device capable of automatically inputting, storing, manipulating, and outputting information in virtue of inputting, storing, manipulating, and outputting representations of that information," where "[t]hese information processes occur in accordance with a finite set of rules that are effective and that are, in some sense, in the machine itself." Another proposal, due to Copeland (1996) and specifically intended to include functions that a Turing machine cannot compute, is that a computer is a device capable of solving a class of problems, that can represent both the problems and their solution, contains some number of primitive operations (some of which may not be Turing computable), can sequence

these operations in some predetermined way, and has a provision for feedback (see Turing, Alan).

To the question of whether the mind/brain is a connectionist versus a classical computer, discussion has centered around Fodor and Pylyshyn's (1988) argument in favor of classicism. In their view, the claim that the mind/brain is a connectionist computer should be rejected on the grounds of the premises that

1. cognitive capacities exhibit *systematicity* (where 'systematicity' is the fact that some capacities are intrinsically connected to others—e.g., if a native speaker of English knows how to say "John loves the girl," the speaker will know how to say "The girl loves John") and
2. this feature of systematicity cannot be explained by reference to connectionist models (unless they are implementations of classical models), whereas
3. it can be explained by reference to classical models.

Fodor and Pylyshyn's reasoning is that it is "characteristic" of classical systems but not of connectionist systems to be "symbol processors," that is, systems that posit mental representations with constituent structure and then process these representations in a way that is sensitive to that structure. Given such features, classical models can explain systematicity; but without such features, as in connectionist machines, it is a mystery. Fodor and Pylyshyn's challenge has generated a host of responses, from both philosophers and computer scientists. In addition, the argument itself has evolved in two significant ways. First, it has been emphasized that to be a counterexample to premise 2, a connectionist model must not only exhibit systematicity, it must also *explain* it. Second, it has been claimed that what needs explaining is not just that human cognitive capacities are systematic; it is that they are *necessarily* so (on the basis of scientific law).

The critical points regarding premise 3 are especially important. The force of Fodor and Pylyshyn's argument rests on classical systems being able to do something that connectionist systems (at a cognitive, nonimplementational level) cannot. However, it has been pointed out that the mere fact that a system is a symbol processer (and hence classical) does not *by itself* explain systematicity, much less the lawful necessity of it; additional assumptions must be made about the system's computational resources. Thus, if the critics are right that connectionist systems of *specific* sorts can also explain systematicity, then the explanatory asymmetry between classicism and connectionism will no longer hold (that is, neither the fact that a system is classical nor the fact that it is connectionist can, taken by itself, explain systematicity, while either fact can explain systematicity when appropriately supplemented), and Fodor and Pylyshyn's argument will be unsound.

The Representational Assumption

Following Peirce (Hartshorne, Weiss, and Burks 1931–58), one can say that any representation has four aspects essential to its being a representation: (i) It is realized by a representation bearer; (ii) it has *semantic content* of some sort; (iii) its semantic content is *grounded* somehow; and (iv) it has *significance* for (that is, it can function as a representation for) the person who has it. (Peirce's terminology was somewhat different. He spoke of a representation's bearer as the "material qualities" of the representation and the content as the representation's "object.")

In Peirce, a representation has significance for a person insofar as it produces a certain effect or "interpretant." This conception of representation is extremely useful for exploring the view of cognitive science vis-à-vis mental representation, for it leads one to ask: What is the representation bearer, semantic content, and ground of *mental* representation, and how do mental representations have significance for the people in whose mind/brains they reside?

A **representation bearer** is an entity or state that has semantic properties considered with respect to its nonrepresentational properties. The representation bearers for mental representations, according to cognitive science, are computational structures or states. If the mind/brain is a classical computer, its representation bearers are data structures; if it is a connectionist computer, its *explicit* representation bearers are patterns of activation of nodes in a network. It is also claimed that connectionist computers *implicitly* represent by means of their connection weights.

Much of the theoretical work of cognitive psychologists consists of claims regarding the content of the representations used in exercising one or another cognitive capacity. For example, psycholinguists posit that in understanding a sentence, people unconsciously form representations of the sentence's phonological structure, words, syntactic structure, and meaning. Although psychologists do not know enough about mental representation as a system to theorize about its semantics in the sense in which linguistics provides the formal

semantics of natural language, if one reflects on what it is that mental representations are hypothesized to explain—certain features of cognitive capacities—one can plausibly infer that the semantics of human mental representation systems must have certain characteristics. In particular, a case can be made that people must be able to represent (i) specific objects; (ii) many different kinds of objects including concrete objects, sets, properties, events and states of affairs in the world, in possible worlds, and in fictional worlds, as well as abstract objects such as universals and numbers; (iii) both objects (*tout court*) and aspects of those objects (or something like extension and intension); and (iv) both correctly and incorrectly.

Although cognitive psychologists have concerned themselves primarily with the representation bearers and semantic content of mental representations, the hope is that eventually there will be an account of how the computational states and structures that function as representation bearers come by their content. In virtue of what, for example, do lexical representations represent specific words? What makes an edge detector represent edges? Such theories are sometimes described as one or another form of ''semantics'' (e.g., informational semantics, functional role semantics), but this facilitates a confusion between theories of content and theories of what determines that content. A preferable term is *theory of content determination*. Philosophers have been concerned both (a) to delineate the basic relation between a representation's having a certain content and its having a certain ground and (b) to sketch alternative theories of content determination, that is, theories of what that ground might be. A common view is that the basic relation is strong supervenience, as defined by Kim (1984). Others, such as Poland (1994), believe that a stronger relation of *realization* is required.

How is the content of mental representations determined? Proposals appeal either exclusively or in some combination to the structure of the representation bearer (Palmer 1978); actual historical (Devitt 1981) or counterfactual causal relations (Fodor 1987) between the representation bearer and phenomena in the world; actual and counterfactual (causal, computational, inferential) relations between the representation-bearer state and other states in the mind/brain (Harman 1987; Block 1987); or the information-carrying or other *functions* of the representation-bearer state and associated components (based on what they were selected for in evolution or learning) (Millikan 1984; Papineau 1987).

Arguably, any adequate theory of content determination will be able to account for the full range of semantic properties of mental representational systems. On this criterion, all current theories of content determination are inadequate. For example, theories that ground representational content in an isomorphism between some aspect of a representation (usually, either its formal structure or its functional role) and what it represents do not seem to be able to explain how one can represent *specific* objects, such as a favorite coffee mug. In contrast, theories that rely on actual, historical causal relations can easily explain the representation of specific objects but do not seem to be able to explain the representation of sets or kinds of objects (e.g., all coffee mugs). It is precisely because single-factor theories of content determination do not seem to have the resources to explain all aspects of people's representational capacities that many philosophers have turned to two-factor theories, such as theories that combine causal relations with function (so-called teleofunctional theories) and theories that combine causal relations with functional role (so-called two-factor theories).

The fourth aspect of a mental representation, in the Peircian view, is that the content of a representation must have significance *for* the representer. In the information-processing paradigm, this amounts to the fact that for each representation, there will be a set of computational consequences of that representation being ''entertained'' or ''activated,'' and in particular a set of computational consequences that are *appropriate* given the content of the representation (von Eckardt 1993, chap. 8).

The Viability of Cognitive Science

Cognitive science has been criticized on several grounds. It has been claimed that there are phenomena within its domain that it does not have the conceptual resources to explain; that one or more of its foundational assumptions are problematic and, hence, that the research program grounded in those assumptions can never succeed; and that it can, in principle, be eliminated in favor of pure neuroscience.

The list of mental phenomena that, according to critics, cognitive science will never be able to explain include that of people ''making sense'' in their actions and speech, of their having sensations, emotions, and moods, a self and a sense of self, consciousness, and the capabilities of insight and creativity and of interacting closely and directly with their physical and social environments. Although no impossibility proofs have been offered, when

cognitive science consisted simply of a top-down, "symbolic" computational approach, this explanatory challenge had a fair amount of intuitive plausibility. However, given the increasing importance of neuroscience within cognitive science and the continuing evolution of the field, it is much less clear today that no cognitive science explanation of such phenomena will be forthcoming. The biggest challenge seems to be the "explanatory gap" posed by phenomenal consciousness (Levine 1983) (see Consciousness).

The major challenge against the computational assumption is the claim that the notion of a computer employed within cognitive science is vacuous. Specifically, Putnam (1988) has offered a proof that every ordinary open system realizes every abstract finite automaton, and Searle (1990) has claimed that implementation is not an objective relation and, hence, that any given system can be seen to implement any computation if interpreted appropriately. Both claims have been disputed. For example, it has been pointed out that Putnam's result relies on employing both an inappropriate computational formalism (the formalism of finite state automata, which have only *monadic* states), an inappropriately weak notion of implementation (one that does not require the mapping from computational to physical states to satisfy *counterfactual* or lawful state–state transitions), and an inappropriately permissive notion of a physical state (one that allows for rigged disjunctions). When these parameters of the problem are made more restrictive, implementations are much harder to come by (Chalmers 1996). However, in defense of Putnam, it has been suggested that this response does not get to the heart of Putnam's challenge, which is to develop a theory that provides necessary and sufficient criteria to determine whether a class of computations is implemented by a class of physical systems (described at a given level). An alternative approach to implementation might be based on the notion of realization of a (mathematical) function by a "digital system" (Scheutz 1999).

The major challenge to the representational assumption is the claim that the project of finding an adequate theory of content determination is doomed to failure. One view is that the attempt by cognitive science to explain the intentionality of cognition by positing mental representations is fundamentally confused (Horst 1996). The argument is basically this: According to cognitive science, a mental state, as ordinarily construed, has some content property in virtue of the fact that it is identical to (supervenes on) a representational state that has some associated semantic property. There are four ways of making sense of this posited semantic property. The semantic property in question is identical to one of the following options:

1. the original content property,
2. an interpreter-dependent semiotic-semantic property,
3. a pure semiotic-semantic property of the sort posited in linguistics, or
4. some new theoretical "naturalized" property.

However, according to Horst (1996), none of these options will work. Options 1 and 2 are circular and hence uninformative. Option 3 is ruled out on the grounds that there is no reason to believe that there are such properties. And option 4 is ruled out on the grounds that naturalized theories of content determination do not have the conceptual resources to deliver the kind of explanation required.

In response, it has been argued that Horst's case against option 3 is inadequate and that his argument against option 4 shows a misunderstanding of what theories of content determination are trying to achieve (von Eckardt 2001). Horst's reason for ruling out option 3 is that he thinks that people who believe in the existence of pure semantic properties or relations do so because there are formal semantic theories that posit such properties and such theories have met with a certain degree of success. But (he argues) when scientists develop models (theories) in science, they do so by a process of abstraction from the phenomena *in vivo* that they wish to characterize. Such abstractions can be viewed either as models *of* the real-world phenomena they were abstracted from or as descriptions of the mathematical properties of those phenomena. Neither view provides a license for new ontological claims. Thus, Horst's argument rests on a nonstandard conception of both theory construction (as nothing but abstraction) and models or theories (as purely mathematical). If, *contra* Horst, a more standard conception is substituted, then linguists have as much right to posit pure semantic properties as physicists have to posit strange subatomic particles. Furthermore, the existence of the posited entities will depend completely on how successful they are epistemically.

Horst's case against option 4, again, exhibits a misunderstanding of the cognitive science project. He assumes that the naturalistic ground N of the content of a mental representation C will be such that the truth of the statement "X is N" conjoined with the necessary truths of logic and mathematics will be *logically sufficient* for the truth of the statement "X has C" and that, further, this entailment will be epistemically transparent. He then argues

against there being good prospects for a naturalistic theory of content on the grounds that naturalistic discourse does not have the conceptual resources to build a naturalistic theory that will entail, in an epistemically transparent way, the truths about intentionality. However, as von Eckardt (2001) points out, Horst's conception of naturalization is much stronger than what most current theory of content determination theorists have in mind, *viz.*, strong supervenience or realization (see Supervenience). As a consequence, his arguments that naturalization is implausible given the conceptual resources of naturalistic discourse are seriously misguided.

BARBARA VON ECKARDT

References

Block, N. (1987), "Functional Role and Truth Conditions," *Proceedings of the Aristotelian Society* 61: 157–181.

Chalmers, D. (1996), "Does a Rock Implement Every Finite-State Automaton?" *Synthese* 108: 309–333.

Chomsky, N., and J. Katz (1974), "What the Linguist Is Talking About," *Journal of Philosophy* 71: 347–367.

Copeland, B. (1996), "What Is Computation?" *Synthese* 108: 336–359.

Devitt, M. (1981), *Designation*. New York: Columbia University Press.

Fodor, J. (1987), *Psychosemantics*. Cambridge, MA: MIT Press.

Fodor, J., and Z. W. Pylyshyn (1988), "Connectionism and Cognitive Architecture: A Critical Analysis," in S. Pinker and J. Miller (eds.), *Connections and Symbols*. Cambridge, MA: MIT Press, 1–71.

Franks, B. (1995), "On Explanation in the Cognitive Sciences: Competence, Idealization and the Failure of the Classical Cascade," *British Journal for the Philosophy of Science* 46: 475–502.

Gardner, H. (1985), *The Mind's New Science: A History of the Cognitive Revolution*. New York: Basic Books.

Harman, G. (1987), "(Non–Solipsistic) Conceptual Role Semantics," in E. Lepore (ed.), *New Directions in Semantics*. London: Academic Press.

Hartshorne, C., P. Weiss, and A. Burks (eds.) (1931–58), *Collected Papers of Charles Sanders Peirce*. Cambridge, MA: Harvard University Press.

Horst, S. W. (1996), *Symbols, Computation, and Intentionality*. Berkeley and Los Angeles: University of California Press.

Kim, J. (1984), "Concepts of Supervenience," *Philosophical and Phenomenological Research* 45: 153–176.

Kuhn, T. (1970), *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Levine, J. (1983), "Materialism and Qualia: The Explanatory Gap," *Pacific Philosophical Quarterly* 64: 354–361.

Marr, D. (1982), *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W. H. Freeman.

Mathews, R. (1991), "Psychological Reality of Grammars," in A. Kasher (ed.), *The Chomskyan Turn*. Oxford and Cambridge, MA: Basil Blackwell.

Millikan, R. (1984), *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Palmer, S. (1978), "Fundamental Aspects of Cognitive Representation," in E. Rosch and B. Lloyd (eds.), *Cognition and Categorization*. Hillsdale, NJ: Erlbaum.

Papineau, D. (1987), *Reality and Representation*. Oxford: Blackwell.

Poland, J. (1994), *Physicalism: The Philosophical Foundations*. Oxford: Oxford University Press.

Putnam, H. (1988), *Representation and Reality*. Cambridge, MA: MIT Press.

Scheutz, M. (1999), "When Physical Systems Realize Functions," *Minds and Machines* 9: 161–196.

Searle, J. (1990), "Is the Brain a Digital Computer?" *Proceedings and Addresses of the American Philosophical Association* 64: 21–37.

Simon, H. (1981), "Cognitive Science: The Newest Science of the Artificial," in D. A. Norman (ed.), *Perspectives on Cognitive Science*. Norwood, NJ: Ablex Publishing, 13–26.

Soames, S. (1984), "Linguistics and Psychology," *Linguistics and Philosophy* 7: 155–179.

von Eckardt, B. (2001), "In Defense of Mental Representation," in P. Gardenfors, K. Kijania–Placek, and J. Wolenski (eds.), *Proceedings of the 11th International Congress of Logic, Methodology and Philosophy of Science*. Dordrecht, Netherlands: Kluwer.

——— (1993), *What Is Cognitive Science?* Cambridge, MA: MIT Press.

See also **Consciousness; Physicalism; Psychology, Philosophy of; Supervenience**

COGNITIVE SIGNIFICANCE

One of the main objectives of logical empiricism was to develop a formal criterion by which cognitively significant statements, which are true or false, could be delineated from meaningless ones, which are neither. The desired criterion would specify, and in some way justify, the logical empiricists' conviction that scientific statements were exemplars of significance and metaphysical ones

were decidedly not (see Logical Empiricism). Finding such a criterion was crucial to logical empiricism. Without it there seemed to be no defensible way to distinguish metaphysics from science and, consequently, no defensible way to exclude metaphysics from subjects that deserved serious philosophical attention (see Demarcation, Problem of). Accordingly, several logical empiricists devoted attention to developing a criterion of cognitive significance, including Carnap, Schlick, Ayer, Hempel, and, to a lesser degree, Reichenbach.

Scientific developments also motivated the project in two related ways. First, physics and biology were demonstrating that *a priori* metaphysical speculations about empirical matters were usually erroneous and methodologically misguided. Hans Driesch's idea of an essential entelechy was no longer considered scientifically respectable, and the intuitive appeal of the concept of absolute simultaneity was shown to be misleading by Einstein (Feigl 1969). Scientific results demonstrated both the necessity and the fruitfulness of replacing intuitive convictions with precise, empirically testable hypotheses, and logical empiricists thought the same methodology should be applied to philosophy. Formulating a defensible criterion that ensured the privileged epistemological status of science, and revealed the vacuity of metaphysics, was thought crucial to the progress and respectability of philosophy.

Second, many scientific discoveries and emerging research programs, especially in theoretical physics, were considerably removed from everyday observable experience and involved abstract, mathematically sophisticated theories. The logical empiricists felt there was a need for a formal systematization of science that could clarify theoretical concepts, their interrelations, and their connection with observation. The emerging tools of modern mathematical logic made this task seem imminently attainable. With the desire for clarity came the pursuit of a criterion that could sharply distinguish these scientific developments, which provided insights about the world and constituted advances in knowledge, from the obfuscations of metaphysics.

Formulation of a cognitive significance criterion requires an empirical significance criterion to delineate empirical from nonempirical statements and a criterion of analyticity to delineate analytic from synthetic statements (see Analyticity). Most logical empiricists thought analytically true and false statements were meaningful, and most metaphysicians thought their claims were true but not analytically so. In their search for a cognitive significance criterion, as the principal weapon of their antimetaphysical agenda, the logical empiricists focused on empirical significance.

The Verifiability Requirement

The first attempts to develop the antimetaphysical ideas of the logical empiricists into a more rigorous criterion of meaningfulness were based on the verifiability theory of meaning (see Verifiability). Though of auxiliary importance to the rational reconstruction in the *Aufbau*, Carnap (1928a) claimed that a statement was verifiable and thereby meaningful if and only if it could be translated into a constructional system; for instance, by reducing it (at least in principle) to a system about basic physical objects or elementary experiences (§179) (see Carnap, Rudolf). Meaningful questions have verifiable answers; questions that fail this requirement are pseudo-questions devoid of cognitive content (§180).

The first explicit, semiformal criterion originated with Carnap in 1928. With the intention of demonstrating that the realism/idealism debate, and many other philosophical controversies, were devoid of cognitive significance, Carnap (1928b) presented a criterion of **factual content**:

> If a statement p expresses the content of an experience E, and if the statement q is either the same as p or can be derived from p and prior experiences, either through deductive or inductive arguments, then we say that q is 'supported by' the experience E . . . . A statement p is said to have 'factual content' if experiences which would support p or the contradictory of p are at least conceivable, and if their characteristics can be indicated.
>
> (Carnap 1967, 327)

Only statements with factual content are empirically meaningful. Notice that a fairly precise inferential method is specified and that a statement has factual content if there are conceivable experiments that could support it. Thus, the earliest formal significance criterion already emphasized that possible, not necessarily actual, connection to experience made statements meaningful.

Carnap ([1932] 1959) made three significant changes to his proposal. First, building on an earlier example (1928b, §7), he developed in more detail the role of syntax in determining the meaningfulness of words and statements in natural languages. The ''elementary'' sentence form for a word is the simplest in which it can occur. For Carnap, a word had to have a fixed mode of occurrence in its elementary sentence form to be significant. Besides failing to connect with experience in some way, statements could also be meaningless because they

contained sequences of words that violated the language's syntactic rules, or its "logical syntax." According to Carnap, "Dog boat in of" is meaningless because it violates grammatical syntax, and "Our president is definitely a finite ordinal," is meaningless because it violates logical syntax, 'president' and 'finite ordinal' being members of different logical categories. The focus on syntax led Carnap to contextualize claims of significance to specific languages. Two languages that differ in syntax differ in whether words and word sequences are meaningful.

Second, Carnap ([1932] 1959) no longer required statements to be meaningful by expressing conceivable states of affairs. Rather, statements are meaningful because they exhibit appropriate deducibility relations with protocol statements whose significance was taken as primitive and incorrigible by Carnap at that time (see Protocol Sentences). Third, Carnap did not specify exactly how significant statements must connect to protocol statements, as he had earlier (1928b). In 1932, Carnap would ascertain a word's meaning by considering the elementary sentence in which it occurred and determining what statements entailed and were entailed by it, the truth conditions of the statement, or how it was verified—considerations Carnap then thought were equivalent. The relations were probably left unspecified because Carnap came to appreciate how difficult it was to formalize the significance criterion, and realized that his earlier criterion was seriously flawed, as was shown of Ayer's first formal criterion (see below).

In contrast to antimetaphysical positions that evaluated metaphysical statements as false, Carnap believed his criterion justified a radical elimination of metaphysics as a vacuous enterprise. The defensibility of this claim depended upon the status of the criterion—whether it was an empirical hypothesis that had to be supported by evidence or a definition that had to be justified on other grounds. Carnap ([1932] 1959) did not address this issue, though he labels the criterion as a stipulation. Whether this stipulation was defensible in relation to other possible criteria or whether the statement of the criterion satisfied the criterion itself were questions left unanswered.

In his popularization of the work of Carnap ([1932] 1959) and Schlick ([1932] 1979), Ayer (1934) addressed these questions and stated that a significance criterion should not be taken as an empirical claim about the linguistic habits of the class of people who use the 'meaning' of a word (see Ayer, Alfred Jules; Schlick, Moritz). Rather, it is a different kind of empirical proposition, which, though conventional, has to satisfy an adequacy condition. The criterion is empirical because, to be adequate, it must classify "propositions which by universal agreement are given as significant" as significant, and propositions that are universally agreed to be nonsignificant as nonsignificant (Ayer 1934, 345).

Ayer developed two formalizations of the criterion. The first edition of *Language, Truth, and Logic* contained the proposal that "a statement is verifiable ... if some observation-statement can be deduced from it in conjunction with certain other premises, without being deducible from those other premises alone," where an **observation statement** is one that records any actual or possible observation (Ayer 1946, 11).

Following criticisms (see the following section) of his earlier work, a decade later Ayer (1946) proposed a more sophisticated criterion by distinguishing between directly verifiable statements and indirectly verifiable ones. In conjunction with a set of observation statements, *directly* verifiable statements entail at least one observation statement that does not follow from the set alone. *Indirectly* verifiable statements satisfy two requirements: (1) In conjunction with a set of premises, they entail at least one directly verifiable statement that does not follow from the set alone; and (2) the premises can include only statements that are either analytic or directly verifiable or can be indirectly verified on independent grounds. Nonanalytic statements that are directly or indirectly verifiable are meaningful, whereas analytic statements are meaningful but do not assert anything about the world.

Early Criticisms of the Verifiability Criterion

The verifiability criterion faced several criticisms, which took two general forms. The first, already mentioned in the last section, questioned its status—specifically, whether the statement of the criterion satisfies the criterion. The second questioned its adequacy: Does the criterion ensure that obviously meaningful statements, especially scientific ones, are labeled as meaningful and that obviously meaningless statements are labeled as meaningless?

Criticisms of the first form often mistook the point of the criterion, construing it as a simple empirical hypothesis about how the concept of meaning is understood or a dogmatic stipulation about how it should be understood (Stace 1935). As mentioned earlier, Ayer (1934, 1946) clearly recognized that it was not this type of empirical claim, nor was it an arbitrary definition. Rather, as Hempel (1950) later made clear, the criterion was intended to clarify

and explicate the idea of a meaningful statement. As an explication it must accord with intuitions about the meaningfulness of common statements and suggest a framework for understanding how theoretical terms of science are significant (see Explication). The metaphysician can deny the adequacy of this explication but must then develop a more liberal criterion that classifies metaphysical claims as significant while evaluating clearly meaningless assertions as meaningless (Ayer 1934).

Criticisms of the second form often involved misinterpretations of the details of the criterion, due partially to the ambiguity of what was meant by 'verifiability.' For example, in a criticism of Ayer (1934), Stace (1935) argued that the verifiability criterion made all statements about the past meaningless, since it was in principle impossible to access the past and therefore verify them. His argument involved two misconceptions. First, Stace construed the criterion to require the possibility of conclusive verification, for instance a complete reduction of any statement to (possible) observations that could be directly verified. Ayer (1934) did not address this issue, but Schlick ([1932] 1979), from whose work Ayer drew substantially, emphasized that many meaningful propositions, such as those concerning physical objects, could never be verified conclusively. Accepting Neurath's criticisms in the early 1930s, Carnap accepted that no statement, including no protocol statement, was conclusively verified (see Neurath, Otto). Recall also that Carnap (1928b) classified statements that were "supported by" conceivable experiences—not conclusively verified—as meaningful.

Second, Stace's argument depended on the ambiguity of "possible verification," which made early formulations of the criterion misleadingly unclear (Lewis 1934). The possibility of verification can have three senses: practical possibility, empirical possibility, and logical possibility. Practical possibility was not the intended sense: "There are 10,000-foot mountains on the moon's far side" was meaningful in the 1930s, though its verification was practically impossible (Schlick [1932] 1979).

However, Carnap (1928a, 1928b, [1932] 1959), Schlick ([1932] 1979), and Ayer (1934) were silent on whether empirical or logical possibility divided the verifiable from the unverifiable. Stace thought time travel was empirically impossible. The question was therefore whether statements about past events were meaningful for which no present evidence was available, and no future evidence would be.

In the first detailed analysis of the verifiability criterion, Schlick ([1936] 1979) stated that the logical impossibility of verification renders a statement nonsignificant. Empirical impossibility, which Schlick understood as contradictng the "laws of nature," does not entail nonverifiability. If it did, Schlick argued, the meaningfulness of a putative statement could be established only by empirical inquiry about the laws of nature. For Schlick, this conflated a statement's meaning with its truth. The meaning of a statement is determined ("bestowed") by logical syntax, and only with meaning fixed *a priori* can its truth or falsity be assessed. Furthermore, since some lawlike generalizations are yet to be identified and lawlike generalizations are never established with absolute certainty, it seems that a sharp boundary between the empirically impossible and possible could never be determined. Hence, there would be no sharp distinction between the verifiable and unverifiable, which Schlick found unacceptable.

For Schlick ([1936] 1979), questions formulated according to the rules of logical grammar are meaningful if and only if it is logically possible to verify their answers. A state of affairs is logically possible for Schlick if the statement that describes it conforms to the logical grammar of language. Hence, meaningful questions may concern states of affairs that contradict well-supported lawlike generalizations. Schlick's position implies that the set of meaningful questions is an extension of the set of questions for which verifiable answers can be imagined. Questions about velocities greater than light are meaningful according to Schlick, but imagining how they could be verified surpasses our mental capabilities.

Schlick's emphasis on logical possibility was problematic because it was unclear that the verification conditions of most metaphysical statements are, or entail, logical impossibilities. In contrast, Carnap ([1936–7] 1965) and Reichenbach (1938) claimed that metaphysical statements were nonsignificant because no *empirically* possible process of confirmation could be specified for them (see Reichenbach, Hans). Furthermore, if only the *logical* possibility of verification were required for significance, then the nonsignificance of metaphysical statements could no longer be demonstrated by demanding an elucidation of the circumstances in which they could be verified. Metaphysicians can legitimately respond that such circumstances may be difficult or impossible to conceive because they are not empirically possible. Nevertheless, the circumstances may be logically possible, and hence the metaphysical statements may be significant according to Schlick's position.

Faced with the problematic vagueness of the early criteria, a formal specification of the criterion was thought to be crucial. Berlin (1939) pointed out

that the early verifiability criteria were open to objections from metaphysicians because the details of the experiential relevance required of meaningful statements were left unclear: ''Relevance is not a precise logical category, and fantastic metaphysical systems may choose to claim that observation data are 'relevant' to their truth'' (233).

With formalizations of the criterion, however, came more definitive criticisms. Ayer's (1946, 39) first proposal was seriously flawed because it seemed to make almost all statements verifiable. For any grammatical statement S—for instance ''The Absolute is peevish''—any observation statement O, and the conditional $S \rightarrow O$, S and $S \rightarrow O$ jointly entail O, though neither of them alone usually does. According to Ayer's criterion, therefore, S and $S \rightarrow O$ are meaningful except in the rare case that $S \rightarrow O$ entails O (Berlin 1939).

Church (1949) presented a decisive criticism of Ayer's (1946, 13) second proposal. Consider three logically independent observation statements O_1, O_2, O_3 and any statement S. The disjunction $(\neg O_1 \wedge O_2) \vee (\neg S \wedge O_3)$ is directly verifiable, since in conjunction with O_1 it entails O_3. Also, $(\neg O_1 \wedge O_2) \vee (\neg S \wedge O_3)$ and S together entail O_2. Hence, by Ayer's criterion, S is indirectly verifiable, unless $(\neg O_1 \wedge O_2) \vee (\neg S \wedge O_3)$ alone entails O_2, which implies $\neg S$ and O_3 entail O_2 so that $\neg S$ is directly verifiable. Thus, according to Ayer's criterion, any statement is indirectly verifiable, and therefore significant, or its negation is directly verifiable, and thereby meaningful.

Nidditch (1961) pointed out that Ayer's (1946) proposal could be amended to avoid Church's (1949) criticism by specifying that the premises could only be analytic, directly verifiable, or indirectly verifiable on independent grounds *and* could only be composed of such statements. Thus that $(\neg O_1 \wedge O_2) \vee (\neg S \wedge O_3)$ and S together entail O_2 does not show that S is indirectly verifiable because $(\neg O_1 \wedge O_2) \vee (\neg S \wedge O_3)$ contains a statement (S) that has not been shown to be analytic, directly verifiable, or independently verifiable on independent grounds. Unfortunately, Scheffler (1963) pointed out that according to Nidditch's (1961) revised criterion, an argument similar to Church's (1949) with the disjunction $\neg O_2 \vee (S \wedge O_1)$ shows that any statement S is significant, unless it is a logical consequence of an observation statement. Scheffler (1963) also pointed out that Ayer's second proposal makes any statement of the form $S \wedge (O_1 \rightarrow O_2)$ significant, where O_1, O_2 are logically independent observation sentences and S is any statement. $S \wedge (O_1 \rightarrow O_2)$ entails $O_2$when conjoined with O_1 and neither the conjunction nor O_1 entails O_2 alone.

Beyond Verifiability: Carnap and Hempel

While Ayer first attempted to formalize the verifiability criterion, Carnap ([1936–7] 1965) recognized the obvious weaknesses of verifiability-based significance criteria. At roughly the same time, in the light of Tarski's rigorous semantic account of truth, Carnap was coming to accept that a systematic (that is, nonpragmatic) account might be possible for other concepts, such as 'confirmation.' He subsequently refocused the question of cognitive significance away from verifiability, which seemed to connote the possibility of definitive establishment of truth, to confirmability—the possibility of obtaining evidence, however partial, for a statement. In particular, Carnap thought a justifiable significance criterion could be formulated if an adequate account of the confirmation of theory by observation were available. A better understanding of the latter would provide a clearer grasp of how scientific terms are significant due to their connection to observation and prediction and how metaphysical concepts are not, because they lack this connection. Yet, insights into the nature of confirmation of theory by observation do not alone determine the form of an adequate significance criterion. Rather, these insights were important because Carnap ([1936–7] 1965) radically changed the nature of the debate over cognitive significance.

Carnap reemphasized (from his work in 1932) that what expressions are cognitively significant depends upon the structure of language, and hence a criterion could be proposed relative to only a specific language. He distinguished two kinds of questions about cognitive significance: those concerning ''historically given language system[s]'' and those concerning constructible ones (Carnap [1932] 1959, 237). Answers to the two kinds of questions are evaluated by different standards. To be meaningful in the first case, an expression E must be a sentence of L, which is determined by the language's syntax, and it must ''fulfill the empiricist criterion of meaning'' (167) for L. Carnap does not disclose the exact form of the criterion—verifiability, testability, or confirmability—for a particular language, such as English.

The reason for Carnap's silence, however, was his belief that the second type of question posed a more fruitful direction for the debate. The second type of question is practical, and the answers are proposals, not assertions. Carnap ([1936–7] 1965) remarked that he was no longer concerned to argue directly that metaphysical statements are not cognitively significant (236). Rather, his strategy was to construct a language L in which every

nonanalytic statement was confirmable by some experimental procedure. Given its designed structure, L will clearly indicate how theoretical statements can be confirmed by observational ones, and it will not permit the construction of metaphysical statements. If a language such as L can be constructed that accords with intuitions about the significance of common statements and is sufficient for the purposes of science, then the onus is on the metaphysician to show why metaphysical statements are significant in anything but an emotive or attitude-expressing way.

In a review paper more than a decade later, Hempel (1950) construed Carnap's ([1936–7] 1965) position as proposing a translatability criterion—a sentence is cognitively significant if and only if it is translatable into an empiricist language (see Hempel, Carl). The vocabulary of an empiricist language L contains observational predicates, the customary logical constants, and any expression constructible from these; the sentence formation rules of L are those of *Principia Mathematica*. The problem Carnap ([1936–7] 1965) attempted to rectify was that many theoretical terms of science cannot be defined in L.

Hempel's interpretation, however, slightly misconstrued Carnap's intention. Carnap ([1936–7] 1965) did not try to demonstrate how theoretical terms could be connected to observational ones in order to *assert* translatability as a criterion of cognitive significance. Rather, in accord with the *principle of tolerance* (Carnap 1934) Carnap's project in 1936–7 was to construct an alternative to metaphysically infused language. The features of the language are then evaluated with respect to the purposes of the language user on pragmatic grounds. Although it seems to conflict with his position in 1932, Carnap (1963) clarified that a "neutral attitude toward various forms of language based on the principle that everyone is free to use the language most suited to his purposes, has remained the same throughout my life" (18–19). Carnap ([1936–7] 1965) tried to formulate a replacement for metaphysics, rather than directly repudiate it on empiricist grounds.

Three definitions were important in this regard. The forms presented here are slightly modified from those given by Carnap ([1936–7] 1965):

1. The confirmation of a sentence S is completely reducible to the confirmation of a class of sentences C if S is a consequence of a finite subclass of C.
2. The confirmation of S directly incompletely reduces to the confirmation of C if (a) the confirmation of S is not completely reducible to C and (b) there is an infinite subclass C' of mutually independent sentences of C such that S entails, by substitution alone, each member of C'.
3. The confirmation of a predicate P reduces to the confirmation of a class of predicates Q if the confirmation of every full sentence of P with a particular argument (e.g., $P(a)$, in which a is a constant of the language) is reducible to the confirmation of a consistent set of predicates of Q with the same argument, together with their negations.

With these definitions Carnap ([1936–7] 1965) showed how dispositional predicates (for instance, "is soluble in water") S could be introduced into an empiricist language by means of reduction postulates or finite chains of them. These postulates could take the simple form of a reduction pair:

$$(\forall x)(Wx \rightarrow (Dx \rightarrow Sx));$$
$$(\forall x)(Fx \rightarrow (Rx \rightarrow \neg Sx));$$

in which W, D, F, and R designate observational terms and S is a dispositional predicate. (In the solubility example, $Sx =$ "x is soluble in water"; $Dx =$ "x dissolves in water"; $Wx =$ "x is placed in water"; and R and F are other observational terms.) If $Dx \leftrightarrow \neg Rx$ and $Wx \leftrightarrow Fx$, then the reduction pair is a bilateral reduction sentence:

$$(\forall x)(Wx \rightarrow (Dx \leftrightarrow Sx)).$$

The reduction postulates introduce, but do not explicitly define, terms by specifying their logical relations with observational terms. They also provide confirmation relations between the two types of terms. For instance, the above reduction pair entails that the confirmation of S reduces to that of the confirmation of the set $\{W, D, F, R\}$. Carnap ([1936–7] 1965) defined a sentence or a predicate to be confirmable (following definitions 1–3 above) if its confirmation reduces to that of a class of observable predicates (156–7). Reduction postulates provide such a reduction for disposition terms such as S. The reduction pair does not define S in terms of observational terms. If $\neg Wx$ and $\neg Fx$, then Sx is undetermined. However, the conditions in which S or its negation hold can be extended by adding other reduction postulates to the language. Carnap thought that supplementing an empiricist language to include terms that could be introduced by means of reduction postulates or chains of them (for example, if Wx is introduced by a reduction pair) would adequately translate all theoretical terms of scientific theories.

Although it set a more rigorous standard for the debate, Carnap's ([1936–7] 1965) proposal encountered difficulties. Carnap believed that bilateral reduction sentences were analytic, since all the consequences of individual reduction sentences that contained only observation terms were tautologies. Yet, Hempel (1951) pointed out that two bilateral reduction sentences together sometimes entailed synthetic statements that contained only observation terms. Since the idea that the conjunction of two analytic sentences could entail synthetic statements was counterintuitive, Hempel made the important suggestion that analyticity and cognitive significance must be relativized to a specific language *and* a particular theoretical context. A bilateral reduction sentence could be analytic in one context but synthetic in a different context that contained other reduction postulates.

Hempel (1950) also argued that many theoretical terms, for instance "gravitational potential" or "electric field," could not be translated into an empiricist language with reduction postulates or chains of them. Introducing a term with reduction postulates provides some sufficient and necessary observation conditions for the term, but Hempel claimed that this was possible only in simple cases, such as electric fields of a simple kind. Introducing a theoretical term with reduction sentences also unduly restricted theoretical concepts to observation conditions. The concept of length could not be constructed to describe unobservable intervals, for instance 1×10^{-100} m, and the principles of calculus would not be constructible in such a language (Hempel 1951). Carnap's ([1936–7] 1965) proposal could not accommodate most of scientific theorizing.

Although ultimately untenable, adequacy conditions for a significance criterion were included in Carnap's ([1936–7] 1965) papers, generalized by Hempel (1951) as: If N is a nonsignificant sentence, then all truth-functional compound sentences that nonvacuously contain N must be nonsignificant. It follows that the denial of a nonsignificant sentence is nonsignificant and that a disjunction, conjunction, or conditional containing a nonsignificant component sentence is also nonsignificant. Yet Hempel (1951) was pessimistic that any adequate criterion satisfying this condition and yielding a sharp dichotomy between significance and nonsignificance could be found. Instead, he thought that cognitive significance was a matter of degree:

> Significant systems range from those whose entire extralogical vocabulary consists of observational terms, through theories whose formulation relies heavily on theoretical constructs, on to systems with hardly any bearing on potential empirical findings. (74).

Hempel suggested that it may be more fruitful to compare theoretical systems according to other characteristics, such as clarity, predictive and explanatory power, and simplicity. On these bases, the failings of metaphysical systems would be more clearly manifested.

Of all the logical empiricists' criteria, Carnap's (1956) criterion was the most sophisticated. It attempted to rectify the deficiencies of his 1936–7 work and thereby avoid Hempel's pessimistic conclusions. Scientific languages were divided into two parts, a theoretical language L_T and an observation language L_O. Let V_O be the class of descriptive constants of L_O, and V_T be the class of *primitive* descriptive constants of L_T. Members of V_O designate observable properties and relations such as 'hard,' 'white,' and 'in physical contact with.' The logical structure of L_O contains only an elementary logic, such as a simple first-order predicate calculus.

The descriptive constants of L_T, called theoretical terms, designate unobservable properties and relations such as 'electron' or 'magnetic field.' L_T contains the mathematics required by science along with the "entities" referred to in scientific physical, psychological, and social theories, though Carnap stressed that this way of speaking does not entail any ontological theses. A theory was construed as a finite set of postulates within L_T and represented by the conjunction of its members T. A finite set of correspondence rules, represented by the conjunction of its members C, connects terms of V_T and V_O.

Within this framework Carnap (1956) presented three definitions, reformulated as:

D1. A theoretical term M is *significant relative* to a class K with respect to L_T, L_O, T, and C $=_{\text{df}}$ if (i) $K \subset V_T$, (ii) $M \notin K$, and (iii) there are three sentences S_M, $S_K \in L_T$, and $S_O \in L_O$ such that:
   (a) S_M contains M as the only descriptive term.
   (b) The descriptive terms in S_K belong to K.
   (c) $(S_M \wedge S_K \wedge T \wedge C)$ is consistent.
   (d) $(S_M \wedge S_K \wedge T \wedge C)$ logically implies S_O.
   (e) $\neg[(S_K \wedge T \wedge C)$ logically implies $S_O]$.

D2. A theoretical term M_n is *significant* with respect to L_T, L_O, T, and C $=_{\text{df}}$ if there is a sequence of theoretical constants $<M_1, \ldots M_n>$ $(M_i \in V_T)$ such that every M_i is significant relative to $\{M_1, \ldots M_{i-1}\}$ with respect to L_T, L_O, T, and C.

D3. An expression A of L_T is a significant sentence of $L_T =_{\text{df}}$ if (i) A satisfies the rules of

formation of L_T and (ii) every descriptive term in A is significant, as in D2.

These definitions, especially D1 (d) and (e), are intended to explicate the idea that a significant term must make a predictive difference. Carnap was aware that observation statements can often be deduced only from theoretical statements containing several theoretical terms. With D2 Carnap implicitly distinguishes between theoretical terms whose significance depends on other theoretical terms and those that acquire significance independently of others. In contrast to his work in 1936–7, and in accord with Hempel's (1951) relativization of analyticity and cognitive significance, Carnap (1956) specified that the significance of theoretical terms is relativized to a particular language *and* a particular theory T.

With the adequacy of his proposal in mind, Carnap (1956, 54–6) proved an interesting result. Consider a language in which V_T is divided into empirically meaningful terms V_1 and empirically meaningless terms V_2. Assume that C does not permit any implication relation between those sentences that contain only V_1 or V_O terms and those sentences that contain only V_2 terms. For a given theory T that can be resolved into a class of statements T_1 that contain only terms from V_1, and T_2 that contain only terms of V_2, then a simple but adequate significance criterion can be given. Any theoretical term that occurs only in isolated sentences, which can be omitted from T without affecting the class of sentences of L_O that it entails in conjunction with C, is meaningless.

The problem is that this criterion cannot be utilized for a theory T' equivalent to T that cannot be similarly divided. Carnap (1956), however, showed by indirect proof that his criterion led to the desired conclusion that the terms of V_2 were not significant relative to T' (L_O, L_T, and C) and that therefore the criterion was not too liberal.

The Supposed Failure of Carnap

Kaplan (1975) raised two objections to Carnap's (1956) criterion that were designed to show that it was too liberal and too restrictive. Kaplan's first objection utilized the "deoccamization" of $T \wedge C$. The label is appropriate, since the transformation of $T \wedge C$ into its deoccamization $T' \wedge C'$ involves replacing all instances of some theoretical terms with disjunctions or conjunctions of new terms of the same type: an Occam-unfriendly multiplication of theoretical terms. Kaplan proved that any deductive systematization of L_O by $T \wedge C$ is also established by any of its deoccamizations. This motivates his intuition that deoccamization should preserve the empirical content of a theory and, therefore, not change the significance of its theoretical terms.

The objection is as follows: If any members of V_T are significant with respect to T, C, L_T, and L_O, then there must be at least one M_1 that is significant relative to an empty K (D2). Yet, if $T \wedge C$ is deoccamized such that M_1 is resolved into two new terms M_{11} and M_{12} that are never found apart, then the original argument that satisfied D1 can no longer be used, since $T' \wedge C'$ do not provide similar logical relationships for M_{11} and M_{12} individually. Hence, the sequence of theoretical terms required by D2 will have no first member. Subsequently, no chain of implications that establishes the significance of successive theoretical terms exists. Although deoccamization preserves the deductive systematization of L_O, according to Carnap's criterion it may render every theoretical term of $T' \wedge C'$ meaningless and therefore render $T' \wedge C'$ devoid of empirical content.

Creath (1976) vindicated the core of Carnap's (1956) criterion by generalizing it to accommodate *sets* of terms, reformulated as:

D1′. A theoretical term M is *significant relative* to a class K with respect to L_T, L_O, T, and $C =_{\text{df}}$ if (i) $K \subset V_T$, (ii) $M \notin K$, (iii) there is a class J such that $J \subset V_T$, $M \in J$, but J and K do not share any members, and (iv) there are sentences S_J, $S_K \in L_T$, and $S_O \in L_O$ such that:
- (a) S_J contains members of J as the only descriptive terms.
- (b) The descriptive terms in S_K belong to K.
- (c) $(S_J \wedge S_K \wedge T \wedge C)$ *is* consistent.
- (d) $(S_J \wedge S_K \wedge T \wedge C)$ logically implies S_O.
- (e) $\neg$[$(S_K \wedge T \wedge C)$ logically implies S_O].
- (f) It is not the case that $(\exists J')(J' \subset J)$ and sentences $S_{J'}$, $S_{K'} \in L_T$, and $S_{O'} \in L_O$ such that:
  - (f1) $S_{J'}$ contains only terms of J' as its descriptive terms.
  - (f2) The descriptive terms of $S_{K'}$ belong to K.
  - (f3) $(S_{J'} \wedge S_{K'} \wedge T \wedge C)$ is consistent.
  - (f4) $(S_{J'} \wedge S_{K'} \wedge T \wedge C)$ logically implies $S_{O'}$.
  - (f5) $\neg$[$(S_{K'} \wedge T \wedge C)$ logically implies $S_{O'}$].

D2′. A theoretical term M_n is *significant* with respect to L_T, L_O, T and $C =_{\text{df}}$ if there is a sequence of sets $\langle J_1, \ldots J_n \rangle$ ($M_n \in J_n$ and $J_i \subset V_T$) such that every member of every set J_i is significant relative to the union of

J_1 through J_{i-1} with respect to L_T, L_O, T and C.

Condition (f) ensures that each member of J is required for the significance of the entire set. Creath (1976) points out that any term made significant by D1 and D2 of Carnap (1956) is made significant by D1′ and D2′ and that according to the generalized criterion, Kaplan's (1975) deoccamization criticism no longer holds.

Kaplan (1975) and Rozeboom (1960) revealed an apparent second flaw in Carnap's (1956) proposal: As postulates (definitions for Kaplan's criticism) are added to $T \wedge C$, the theoretical terms it contains may change from cognitively significant to nonsignificant or vice versa. Consider an example from Kaplan (1975) in which $V_O = \{J_O, P_O, R_O\}$; L_O is the class of all sentences of first-order logic with identity that contain no descriptive constants or only those from V_O; $V_T = \{B_T, F_T, G_T, H_T, M_T, N_T\}$; and L_T is the class of all sentences of first-order logic with identity that contain theoretical terms from V_T. Let T be:

$$(T)[(\forall x)(H_T x \rightarrow F_T x)] \wedge [(\forall x)(H_T x \rightarrow (B_T x \vee \neg G_T x))] \wedge [(\forall x)(M_T x \leftrightarrow N_T x)];$$

and let C be:

$$(C)[(\forall x)(R_O x \rightarrow H_T x)] \wedge [(\forall x)(F_T x \rightarrow J_O x)] \wedge [(\forall x)(G_T x \rightarrow P_O x)].$$

G_T, F_T, and H_T are significant with respect to $T \wedge C$ relative to the empty set (see Carnap [1956] D1) and, hence, significant with respect to L_O, L_T, T, and C (see Carnap [1956] D2). R_O is significant relative to $K = \{G_T\}$; M_T and N_T are not significant.

Consider a definitional extension T' of T in an extended vocabulary V_T' and language L_T'. After adding two definitions to T:

$$(\text{DEF1})(\forall x)(D1_T x \leftrightarrow (M_T x \wedge (\exists x) F_T x))$$

and

$$(\text{DEF2})(\forall x)(D2_T x \leftrightarrow (M_T x \rightarrow (\exists x) G_T x)),$$

$D1_T$ is significant relative to the empty set and therefore significant with respect to T', C, L_O, and L_T' (D2). $D2_T$ is significant relative to $K = \{D1_T\}$, and therefore significant with respect to T', C, L_O, and L_T' (D2). M_T, which failed to be significant with respect to T, C, L_O, and L_T, is now significant with respect to T', C, L_O, and L_T'. A similar procedure makes N_T significant. Kaplan thought this showed that Carnap's (1956) criterion was too liberal. The procedure seems able to make any theoretical term significant with respect to some extended language and definition-extended theory, but "definitional extensions are ordinarily thought of as having no more empirical content than the original theory" (Kaplan 1975, 90).

Using the same basic strategy, Rozeboom (1960) demonstrated that extending $T \wedge C$ can transform an empirically significant term into an insignificant one. Consider a term M that is significant with respect to T, C, L_T, and L_O. Rozeboom showed that if postulates (not necessarily definitions) are added to T or to C to form T' or C', in some cases D1(e) will no longer be satisfied, and no other sentences S_M', S_K', S_O' exist by which M could be independently shown to be significant. Furthermore, if $T \wedge C$ is maximally L_O consistent, no theoretical term of L_T is significant, since D1(e) is never satisfied; for any S_O, if $T \wedge C$ is maximally L_O consistent then it alone implies S_O. Rozeboom (1960) took the strength of his criticism to depend upon the claim that for a criterion to be "intuitively acceptable," theoretical terms must retain significance if T or C is extended.

Carnap (1956) can be defended in at least two ways. First, as Kaplan (1975) notes, the criterion was restricted to primitive, nondefined theoretical terms. It was explicitly formulated to avoid criticisms derived from definitional extensions. Defined terms often play an important role in scientific theories, and it could be objected that any adequate criterion should apply directly to theories that contain them. Yet the amendment that any theoretical term within the definiens of a significant defined term must be antecedently shown significant quells these worries (Creath 1976).

Second, Carnap (1956) insisted that terms are significant only *within a particular language* and *for a particular T and C*. He did not intend to formulate a criterion of cognitive significance that held under theory or language change. If Carnap's (1956) work on a significance criterion was an explication of the idea of meaningfulness (Hempel 1950), the explicandum was the idea of a meaningful statement of a particular language in a particular theoretical context, not meaningfulness per se. Hence, Kaplan and Rozeboom's objections, which rely on questionable intuitions about the invariance of significance as $T \wedge C$ changes, are not appropriately directed at Carnap (1956). The fact that Carnap did not attempt such an account is not merely the result of a realization that so many problems would thwart the project. Rather, it is a consequence of the external/internal framework that he believed was the most fruitful approach to the philosophical questions (Carnap 1947).

Furthermore, Rozeboom's acceptability condition is especially counterintuitive, since changes in

T or C designate changes in the connections between theoretical terms themselves or theoretical terms and observation terms. Additional postulates that specify new connections, or changes in the connections, between these terms can obviously change the significance of a theoretical term. Scientific advances are sometimes made when empirical or theoretical discoveries render a theoretical term nonsignificant.

JAMES JUSTUS

References

Ayer, A. J. (1934), "Demonstration of the Impossibility of Metaphysics," *Mind* 43: 335–345.

——— (1946), *Language, Truth, and Logic*, 2nd ed. New York: Dover Publications.

Berlin, I. (1939), "Verification," *Proceedings of the Aristotelian Society* 39: 225–248.

Carnap, R. (1928a), *Der Logische Aufbau der Welt*. Berlin-Schlachtensee: Weltkreis-Verlag.

——— (1928b), *Scheinprobleme in der Philosophie: Das Fremdpsychische und der Realismusstreit*. Berlin-Schlachtensee: Weltkreis-Verlag.

——— ([1932] 1959), "The Elimination of Metaphysics Through Logical Analysis of Language," in A. J. Ayer (ed.), *Logical Positivism*. Glencoe, IL: Free Press, 60–81.

——— (1934), *The Logical Syntax of Language*. London: Routledge Press.

——— ([1936–7] 1965), "Testability and Meaning," in R. R. Ammerman (ed.), *Classics of Analytic Philosophy*. New York: McGraw-Hill, 130–195.

——— (1947), "Empiricism, Semantics, and Ontology," in *Meaning and Necessity*. Chicago: University of Chicago Press, 205–221.

——— (1956), "The Methodological Character of Theoretical Concepts," in H. Feigl and M. Scriven (eds.), *The Foundations of Science and the Concepts of Psychology and Psychoanalysis*. Minneapolis: University of Minnesota Press, 38–76.

——— (1963), "Intellectual Autobiography," in P. Schlipp (ed.), *The Philosophy of Rudolf Carnap*. Peru, IL: Open Court Press, 3–86.

——— (1967), *Logical Structure of the World and Pseudoproblems in Philosophy*. Berkeley and Los Angeles: University of California Press.

Church, A. (1949), "Review of the Second Edition of *Language, Truth and Logic*," *Journal of Symbolic Logic* 14: 52–53.

Creath, R. (1976), "Kaplan on Carnap on Significance," *Philosophical Studies* 30: 393–400.

Feigl, H. (1969), "The Origin and Spirit of Logical Positivism," in P. Achinstein and S. F. Barker (eds.), *The Legacy of Logical Positivism*. Baltimore, MD: Johns Hopkins Press, 3–24.

Hempel, C. (1950), "Problems and Changes in the Empiricist Criterion of Meaning," *Revue Internationale de Philosophie* 11: 41–63.

——— (1951), "The Concept of Cognitive Significance: A Reconsideration," *Proceedings of the American Academy of Arts and Sciences* 80: 61–77.

Kaplan, D. (1975), "Significance and Analyticity," in J. Hintikka (ed.), *Rudolf Carnap, Logical Empiricist*. Dordrecht, Netherlands: D. Reidel Publishing Co., 87–94.

Lewis, C. I. (1934), "Experience and Meaning," *Philosophical Review* 43: 125–146.

Nidditch, P. (1961), "A Defense of Ayer's Verifiability Principle Against Church's Criticism," *Mind* 70: 88–89.

Reichenbach, H. (1938), *Experience and Prediction*. Chicago: University of Chicago Press.

Rozeboom, W. W. (1960), "A Note on Carnap's Meaning Criterion," *Philosophical Studies* 11: 33–38.

Scheffler, I. (1963), *Anatomy of Inquiry*. New York: Alfred A. Knopf.

Schlick, M. ([1932] 1979), "Positivism and Realism," in H. L. Mulder and B. F. B. van de Velde-Schlick (eds.), *Moritz Schlick: Philosophical Papers (1926–1936)*, vol. 2. Dordrecht, Netherlands: D. Reidel Publishing Co., 259–284.

———([1936] 1979), "Meaning and Verification," in H. L. Mulder and B. F. B. van de Velde-Schlick (eds.), *Moritz Schlick: Philosophical Papers (1926–1936)*, vol. 2. Dordrecht, Netherlands: D. Reidel Publishing Co., 456–481.

Stace, W. T. (1935), "Metaphysics and Meaning," *Mind* 44: 417–438.

See also **Analyticity; Ayer, Alfred Jules; Carnap, Rudolf; Corroboration; Demarcation, Problem of; Explication; Feigel, Herbert; Hempel, Carl; Logical Empiricism; Neurath, Otto; Popper, Karl; Rational Reconstruction; Reichenbach, Hans; Schlick, Moritz; Tarski, Alfred; Verifiability; Vienna Circle**

COMPLEMENTARITY

The existence of indivisible interaction quanta is a crucial point that implies the *impossibility of any sharp separation between the behavior of atomic objects and the interaction with the measuring instruments that serve to define the conditions under which the phenomena appear*. In fact, the individuality

of the typical quantum effects finds its proper expression in the circumstance that any attempt at subdividing the phenomena will demand a change in the experimental arrangement, introducing new possibilities of interaction between objects and measuring instruments, which in principle cannot be controlled. Consequently, evidence obtained under different experimental conditions cannot be comprehended within a single picture but must be regarded as **complementary** in the sense that only the totality of the phenomena exhausts the possible information about the objects (Bohr 2000, 209–10).

AU:33

Complementarity is distinctively associated with the Danish physicist Niels Bohr and his attempt to understand the new quantum mechanics (QM) during the heyday of its invention, 1920–1935. Physicists know in practice how to extract very precise and accurate predictions and explanations from QM. Yet, remarkably, even today no one is confident about how to interpret it metaphysically (''M-interpret'' it). That is, there is no compellingly satisfactory account of what sort of objects and relations make up the QM realm, and so, ultimately, no one knows why the precise answers are as they are. In classical mechanics (CM), physicists thought they had a lucidly M-interpretable theory: There was a collection of clearly specified entities, particles, or waves that interacted continuously according to simple laws of force so that the system state was completely specified everywhere and at all times—indeed, specification of just instantaneous position and momenta sufficed. Here the dynamic process specified by the laws of force, expressed in terms of energy and momentum (Bohr's ''causal picture''), generated a uniquely unfolding system configuration expressed in terms of position and time (Bohr's ''space-time picture''). To repeat this for QM, what is needed is a collection of equivalent quantum objects whose interactions and movements in space-time generate the peculiar QM statistical results in ways that are as intuitively clear as they are for CM. (However, even this appealing concept of CM proves too simplistic; there is continuing metaphysical perplexity underlying physics; see e.g., Earman 1986; Hooker 1991, note 13.)

AU:34

That M-interpreting QM is not easy is nicely illustrated by the status of the only agreed-on general ''interpretation'' to which physicists refer, **Born's rule**. It specifies how to extract probabilities from the QM wave function (ψ [''psi'']-function). These are normally associated with particle-like events. But without further M-interpretive support, Born's rule becomes merely part of the recipe for extracting numbers from the QM mathematics. That it does not M-interpret the QM mathematics is made painfully clear by the fact that the obvious conception of the QM state it suggests—a statistical ensemble of particles, each in a definite classical state—is provably not a possible M-interpretation of QM. (For example, no consistent sense can then be made of a superposition of ψ-states, since it is a mathematical theorem that the QM statistics of a superposed state cannot all be deduced from any single product of classical statistical states.)

Bohr does not offer an M-interpretation of QM. He came to think the very idea of such an interpretation incoherent. (In that sense, the term ''Copenhagen interpretation'' is a misnomer; Bohr does not use this label.) Equally, however, Bohr does not eschew all interpretive discussion of QM, as many others do on the (pragmatic or positivist-inspired) basis that confining the use of QM strictly to deriving statistics will avoid error while allowing science to continue. Bohr's position is that this too is profoundly wrong, and ultimately harmful to physics. Instead he offers the doctrine of complementarity as a ''framework'' for understanding the ''epistemological lesson'' of QM (not its ontological lesson) and for applying QM consistently and as meaningfully as possible. (see Folse 1985 for a general introduction. For extensive, more technical analyses, see Faye and Folse 1994; Honner 1987; Murdoch 1987. For one of many critiques, and the opposing Bohrian M-interpretation, see Cushing 1994.)

Although Bohr considered it necessary (''unavoidable'') to continue using the key descriptive concepts of CM, the epistemological lesson of QM was that the basic conditions for their well-defined use were altered by the quantization of QM interactions into discrete unanalyzable units. This, he argued, divided the CM state in two, the conditions for the well-defined use of (1) causal (energy-momentum) concepts and (2) configurational (space-time) concepts being now mutually exclusive, so that only one kind of description could be coherently provided at a time. Both kinds of description were necessary to capture all the aspects of a QM system, but they could not be simply conjoined as in CM. They were now complementary.

Heisenberg's uncertainty relations of QM, such as $\Delta x \Delta p \geq h/2\pi$, where Δx is the uncertainly in the position, Δp is the uncertainty in the momentum, and h is Planck's constant (or more generally the commutation relations, such as $[x, p_x] = ih/2\pi$ where h is again Planck's constant, the magnitude of quantization), are not themselves statements

of complementarity. Rather, they specify the corresponding quantitative relationships between complementary quantities.

These complementary exclusions are implicit in QM ontologies. For instance, single-frequency ("pure") waves must, mathematically, occupy all space, whereas restricting their extent involves superposing waves of different frequencies, with a point-size wave packet requiring use of all frequencies; thus, frequency and position are not uniquely cospecifiable. And since the wavelength λ is the wave velocity V (a constant) divided by the frequency $v(\lambda = V/v)$, uniquely specifying wavelength and uniquely specifying position equally are mutually exclusive. QM associates wavelength with momentum $(p = h/\lambda)$ and energy with frequency $(E = hv)$ in all cases with discrete values and for both radiation (waves) and matter (particles), yielding the QM exclusions. (Note, however, that wave/particle complementarity is but one aspect of causal/configurational complementarity, the aspect concerned with physical conditions that frame coherent superposition versus those that frame localization.)

Precisely why these particular associations (and similar QM associations) should follow from quantization of interaction is not physically obvious, despite Bohr's confident assertion. Of course such associations follow from the QM mathematics, but that presupposes rather than explains complementarity, and Bohr intended complementarity to elucidate QM. The physical and mathematical roots of quantization are still only partially understood. However, it is clear that discontinuity leads to a constraint, in principle, on joint precise specification. Consider initially any quantity that varies with time (t), e.g., energy (E), so that $E = f(t)$.Then across some time interval, $t_1 \rightarrow t_2$, E will change accordingly: $E_1(t_1) \rightarrow E_2(t_2)$.

If E varies continuously, then both E and t are everywhere jointly precisely specifiable because for every intermediate value of t between t_1 and t_2 (say, t_{1+n}) there will be a corresponding value for E: $E_{1+n} = f(t_{1+n})$. Suppose, however, that E (but not t) is quantized, with no allowed value between E_1 and E_2. Then no intermediate E value is available, and energy must remain undefined during at least some part of the transition period. This conclusion can be generalized to any two or more related quantities. The problem is resolved if both quantities are quantized, but there is as yet no satisfactory quantization of space and time (Hooker 1991).

Such inherent mutual exclusions should not be mistaken for merely practical epistemic exclusions (some of Heisenberg's pronouncements notwithstanding). Suppose that the position of an investigated particle i is determined by bouncing ("scattering") another probe particle p off of it, determining the position of i from the intersection of the initial and final momenta of p. However, i will have received an altered momentum in the interaction, and it is tempting to conclude that we are thus excluded from knowing both the position and the momentum of i immediately after the interaction. But in CM the interaction may be retrospectively analyzed to calculate the precise change in momentum introduced by p to i, using conservation of momentum, and so establish both the position and the momentum of i simultaneously. More generally, it is in this manner possible to correct for all measurement interactions and arrive at a complete classical state specified independently of its method of measurement. This cannot, in principle, be done in QM, because of quantization.

Faye (1991) provides a persuasive account of the origins of Bohr's ideas about the applicability of physical concepts in the thought of the Danish philosopher Harald Høffding (a family friend and early mentor of Bohr's) and sets out Bohr's consequent approach. According to Høffding, objective description in principle required a separation between describing a subject and describing a known object (Bohr's "cut" between them) in a way that always permitted the object to be ascribed a unique (Bohr's "unambiguous") spatiotemporal location, state, and causal interaction. These ideas in turn originated in the Kantian doctrine that an objective description of nature requires a well-defined distinction between the knower and the object of knowledge, permitting the unique construction of a well-defined object state, specified in applications of concepts from the synthetic *a priori* (essentially Newtonian) construction of the external world (see Friedman 1992). We have just noted how CM satisfies this requirement.

Contrarily, Bohr insisted, the quantum of action creates an "indissoluble bond" between the measurement apparatus (m-apparatus, including the sentient observer) and the measured (observed) system, preventing the construction of a well-defined system state separate from observing interactions. This vitiates any well-defined, global cut between m-apparatus and system. Creating a set of complementary partial cuts is the best that can now be done. In fact, these circumstances are generalized to all interactions between QM

systems; the lack of a global separation is expressed in their superposition, which defies reduction to any combination of objectively separate states. Consequently, Bohr regarded CM as an idealized physics (achieved, imperfectly, only in the limit $h \to 0$) and QM as a "rational generalization" of it, in the sense of the principle of affinity, the Kantian methodological requirement of continuity.

Bohr's conception of what is required of a physically intelligible theory T can thus be summarized as follows (Hooker 1991, 1994):

BI1. Each descriptive concept A of T has a set of well-defined, epistemically accessible conditions C_A under which it is unambiguously applicable.

BI2. The set of such concepts collectively exhausts, in a complementary way, the epistemically accessible features of the phenomena in the domain of T.

BI3. There is a well-defined, unified, and essentially unique formal structure $S(T)$ that structures and coordinates descriptions of phenomena so that each description is well defined (the various conditions C_A are consistently combined), $S(T)$ is formally complete (Bub 1974), and BI2 is met.

BI4. **Bohr objectivity** (BO) satisfies BI1–3 in the most empirically precise and accurate way available across the widest domain of phenomena while accurately specifying the interactive conditions under which such phenomena are accessible to us.

AU:36

An objective representation of nature thus reflects the interactive access ("point of view") of the knowing subject, which cannot be eliminated. In coming to know nature, we also come to know ourselves as knowers—not fundamentally by being modeled in the theory as *objects* (although this too happens, in part), but by the way the very form of rational generalization reflects our being as knowing *subjects.*

A very different ideal of scientific intelligibility operates in classical physics, and in many proposed M-interpretations of QM. Contrary to BI1, descriptive concepts are taken as straightforwardly characterizing external reality (describing an M, even if it is a strange one). Hence, contrary to BI2, these concepts apply conjointly to describe reality completely. Contrary to BI3 and BI4, an objective theory completely and accurately describes the physical state at each moment in time and provides a unique interactive dynamic history of states for all systems in its domain. Accordingly, measurements are analyzed similarly as the same kinds of dynamic interactions, and statistical descriptions reflect (only) limited information about states and are not fundamental (contrary to common readings of QM). Here the objective representation of nature through invariances eliminates any inherent reference to any subject's point of view. Rather, in coming to know nature we also come to know ourselves as knowers by being modeled in the theory as some *objects* among others so as to remove ourselves from the form of the theory, disappearing as *subjects*. This shift in ideals of intelligibility and objectivity locates the full depth of Bohr's doctrine of complementarity.

AU:37

References

Bohr, N. (2000), "Discussion with Einstein on Epistemological Problems in Atomic Physics," in P. A. Schilpp (ed.), *Albert Einstein: Philosopher-Scientist*, 3rd ed. La Salle, IL: Open Court, 199–242.

AU:38

AU:39

——— (1961), *Atomic Theory and the Description of Nature*. Cambridge: Cambridge University Press.

Bub, J. (1974), *The Interpretation of Quantum Mechanics*. Dordrecht, Netherlands: Reidel.

Cushing, J. T. (1994), "A Bohmian Response to Bohr's Complementarity," in J. Faye and H. J. Folse (eds.), *Niels Bohr and Contemporary Philosophy*. Dordrecht, Netherlands: Kluwer.

Earman, J. (1986), *A Primer on Determinism*. Dordrecht, Netherlands: Reidel.

Faye, J. (1991), *Niels Bohr: His Heritage and Legacy*. Dordrecht, Netherlands: Kluwer.

Faye, J., and H. J. Folse (eds.) (1994), *Niels Bohr and Contemporary Philosophy*. Dordrecht, Netherlands: Kluwer.

Folse, H. J. (1985), *The Philosophy of Niels Bohr*. Amsterdam: North-Holland.

Friedman, M. (1992), *Kant and the Exact Sciences*. Cambridge, MA: Harvard University Press.

Honner, J. (1987), *The Description of Nature*. Oxford: Clarendon.

Hooker, C. A. (1972), "The Nature of Quantum Mechanical Reality: Einstein versus Bohr," in R. G. Colodny (ed.), *Pittsburgh Studies in the Philosophy of Science*, vol. 5. Pittsburgh, PA: University of Pittsburgh Press.

——— (1991), "Physical Intelligibility, Projection, and Objectivity: The Divergent Ideals of Einstein and Bohr," *British Journal for the Philosophy of Science* 42, 491–511.

——— (1994), "Bohr and the Crisis of Empirical Intelligibility: An Essay on the Depth of Bohr's Thought and Our Philosophical Ignorance," in J. Faye and H. J. Folse (eds.), *Niels Bohr and Contemporary Philosophy*. Dordrecht, Netherlands: Kluwer.

Murdoch, D. (1987), *Niels Bohr's Philosophy of Physics*. Cambridge: Cambridge University Press.

C. A. HOOKER

See also **Classical Mechanics; Quantum Measurement Problem; Quantum Mechanics**

COMPLEXITY

See **Unity and Disunity of Science**

CONFIRMATION THEORY

When evidence does not conclusively establish (or refute) a hypothesis or theory, it may nevertheless provide *some* support for (or against) the hypothesis or theory. Confirmation theory is concerned almost exclusively with the latter, where conclusive support (or "countersupport") are limiting cases of confirmation (or disconfirmation). (Included also, of course, is concern for the case in which the evidence is confirmationally irrelevant.) Typically, confirmation theory concerns *potential* support, the impact that evidence *would* have on a hypothesis or theory if learned, where whether the evidence is actually learned or not is not the point; for this reason, confirmation theory is sometimes called the *logic* of confirmation. (For simplicity of exposition for now, theories will be considered separately below and not explicitly mentioned until then.)

It is relevant here to point out the distinction between deductive logic (or deductive evaluation of arguments) and inductive logic (or inductive evaluation of arguments). In **deductive logic**, the question is just *whether or not* the supposed truth of all the premises of an argument gives an *absolute guarantee* of truth to the conclusion of the argument. In **inductive logic**, the question is whether the supposed truth of all the premises of an argument gives *significant support* for the truth of the conclusion, where, ideally, some measure of *to what degree* the premises support the conclusion (which is sometimes called the inductive probability of an argument) would be provided (see Inductive Logic; Induction, Problem of; and Verisimilitude. As in each of these topics also, the question is one of either qualitative or quantitative support that premises or evidence provides to a conclusion or that a hypothesis has. See Carnap, Rudolf, for an idea of degree of confirmation based on his proposed "logical" interpretation of probability and degree of support.) In the theory of the logic of support, confirmation theory is concerned primarily with inductive support, where the theory of deductive support is supposed to be more fully understood.

The concept of confirmation can be divided into a number of subconcepts, corresponding to three distinctions. First, **absolute confirmation** and **incremental confirmation** may be distinguished. In the absolute sense, a hypothesis is confirmed by evidence if the evidence makes (or would make) the hypothesis highly supported; absolute confirmation is about how the evidence "leaves" the hypothesis. In the incremental sense, evidence confirms a hypothesis if the evidence makes the hypothesis *more* highly confirmed (in the absolute sense) than it is (in the absolute sense) without the evidence; incremental confirmation involves a comparison. Second, confirmation can be thought of either *qualitatively* or *quantitatively*. So, in the absolute sense of confirmation, a hypothesis can be, qualitatively, left *more or less* confirmed by evidence, where quantitative confirmation theory attempts to make sense of assigning *numerical degrees* of confirmational support ("inductive probabilities") to hypotheses in light of the evidence. In the incremental sense of confirmation, evidence E may, qualitatively, either confirm, disconfirm, or be evidentially irrelevant to a hypothesis H, where in the quantitative sense, a numerical magnitude (which

can be measured, ''inductive probabilistically,'' in different ways; see below) is assigned to the ''boost'' (positive or negative, if any) that E gives H. Finally, confirmation can be considered to be either **comparative** or **noncomparative**. Noncomparative confirmation concerns just one hypothesis/evidence pair. In comparative confirmation, one can compare how well an E supports an H with how well an E' supports the same H; or one may compare how well an E supports an H with how well the same E supports an H'; or one may compare how well an E supports an H with how well an E' supports an H'.

The exposition below will be divided into two main parts. The first part, ''Nonprobabilistic Approaches,'' will concern different aspects of qualitative confirmation; and the second part, ''Probabilistic Approaches,'' will consider some major quantitative approaches. Almost exclusively, as in the literature, the issue will be incremental confirmation rather than absolute confirmation. Both noncomparative and various kinds of comparative approaches will be described.

Nonprobabilistic Approaches

A simple and natural idea about the confirmation of a general hypothesis of the form ''All *F*s are *G*s'' is that an object confirms the hypothesis if and only if it is both an F and a G (a ''positive instance'' of the hypothesis), disconfirms the hypothesis if and only if it is an F but not a G (a ''negative instance''), and is evidentially irrelevant if and only if it is not even an F (no kind of instance). Hempel ([1945] 1965) calls this **Nicod's criterion** (Nicod 1930). Another natural idea about the confirmation of hypotheses is that if hypotheses H and H' are logically equivalent, then evidence E confirms, disconfirms, or is irrelevant to H if and only if E confirms, disconfirms, or is irrelevant to H', respectively. Hempel calls this the *equivalence condition*, and distinguishes between criteria (definitions or partial definitions) of confirmation and the conditions of adequacy that the criteria should satisfy. Hempel points out that Nicod's criterion does not satisfy the equivalence condition (as long as confirmation, disconfirmation, and evidential irrelevance are mutually exclusive). For example, a hypothesis ''All *F*s are *G*s'' is logically equivalent to ''All non-*G*s are non-*F*s,'' but Nicod's criterion implies that an object that is an F and a G would confirm the former but be irrelevant to the latter, thus violating the equivalence condition. Also, ''All *F*s are *G*s'' is logically equivalent to ''Anything that is both an F and a non-G is both an F and a non-F,'' which Nicod's criterion implies that nothing can confirm (a positive instance would have to be both an F and a non-F).

Thus, Hempel suggests weakening Nicod's criterion. The idea that negative instances disconfirm (i.e., are *sufficient* to disconfirm) is retained. Further, Hempel endorses the **positive-instance criterion**, according to which positive instances are *sufficient* for confirmation. Nicod's criterion can be thought of as containing six parts: necessary and sufficient conditions for all three of confirmation, disconfirmation, and irrelevance. The positive-instance criterion is said to be one-sixth of Nicod's criterion, and it does not lead to the kind of contradiction that Nicod's full criterion does when conjoined with the equivalence condition.

However, the combination of the positive-instance criterion and the equivalence condition (i.e., the proposition that the positive-instance criterion satisfies the equivalence condition) does lead to what Hempel called *paradoxes of confirmation*, also known as **the Ravens paradox** and Hempel's paradox. Hempel's famous example is the hypothesis H: ''All ravens are black.'' Hypothesis H is logically equivalent to hypothesis H': ''All nonblack things are nonravens.'' According to the equivalence condition, anything that confirms H' confirms H. According to the positive-instance criterion, nonblack nonravens (positive instances of H') confirm H'. Examples of nonblack nonravens (positive instances of H') include white shoes, yellow pencils, transparent tumblers, etc. So it follows from the positive-instance criterion plus the equivalence condition that objects of the kinds just listed confirm the hypothesis H that all ravens are black. These conclusions seem incorrect or counterintuitive, and the paradox is that the two seemingly plausible principles, the positive-instance criterion and the equivalence condition, lead, by valid reasoning, to these seemingly implausible conclusions. Further paradoxical consequences can be obtained by noting that the hypothesis H is logically equivalent also to H'', ''All things that are either a raven or not a raven (i.e., all things) are either black or not a raven,'' which has as positive instances any objects that are black and any objects that are not ravens.

Since the equivalence condition is so plausible (if H and H' are logically equivalent, they can be thought of as simply different formulations of the same hypothesis), attention has focused on the positive-instance criterion. Hempel defended the criterion, arguing that the seeming paradoxicalness of the consequences of the criterion is more of a

psychological illusion than a mark of a logical flaw in the criterion:

> In the seemingly paradoxical cases of confirmation, we are often not actually judging the relation of the given evidence E alone to the hypothesis H . . . . [I]nstead, we tacitly introduce a comparison of H with a body of evidence which consists of E in conjunction with additional information that we happen to have at our disposal.
>
> (Hempel [1945] 1965, 19)

So, for example, if one is just given the information that an object is nonblack and a nonraven (where it may happen to be a white shoe or a yellow pencil, but this is not included in the evidence), then the idea is that one should intuitively judge the evidence as confirmatory, ''and the paradoxes vanish'' (20). To assess properly the seemingly paradoxical cases for their significance for the logic of confirmation, one must observe the ''*methodological* fiction'' (as Hempel calls it) that one is in a position to judge the relation between the given evidence *alone* (e.g., that an object is a positive instance of the contrapositive of a universalized conditional) and the hypothesis in question and that there is no other information. This approach has been challenged by some who have argued that confirmation should be thought of as a relation among three things: evidence, hypothesis, and background knowledge (see the section below on Probabilistic Approaches).

Given the equivalence condition, the Ravens paradox considers the question of which of *several* kinds of evidence confirm(s) what one can consider to be a single hypothesis. There is another kind of paradox, or puzzle, that arises in a case of a single body of evidence and multiple hypotheses. In Nelson Goodman's (1965) well-known Grue paradox or puzzle, a 'new' predicate is defined as follows. Let's say that an object A is ''grue'' if and only if *either* (i) A has been observed before a certain time t (which could be now or some time in the future) and A is green *or* (ii) A has not been observed before that time t and A is blue. Consider the hypothesis H that all emeralds are green and the hypothesis H' that all emeralds are grue. And consider the evidence E to be the observation of a vast number of emeralds, all of which have been green. Given that t is now or some time in the future, E is equivalent to E', that the vast number of emeralds observed have all been grue. It is taken that E (the ''same'' as E') confirms H (this is natural enough) but not H'—for in order for H' to be true, exactly all of the unobserved (by t) emeralds would have to be blue, which would seem to be disconfirmed by the evidence. Yet, the evidence E' (or E) consists of positive instances of H'.

Since the positive-instance criterion can be formulated purely syntactically—in terms of simply the logical forms of evidence sentences and hypotheses and the logical relation between their forms—a natural lesson of the Grue example is that confirmation cannot be characterized purely syntactically. (It should be noted that an important feature of Hempel's ([1945] 1965) project was the attempt to characterize confirmation purely syntactically, so that evidence E should, strictly speaking, be construed as evidence statements or sentences, or ''observation reports,'' as he put it, rather than as observations or the objects of observation.) And a natural response to this has been to try to find nonsyntactical features of evidence and hypothesis that differentiate cases in which positive instances confirm and cases in which they do not. And a natural idea here is to distinguish between *predicates* that are ''projectible'' (in Goodman's terminology) and those that are not. Goodman suggested ''entrenchment'' of predicates as the mark of projectibility—where a predicate is entrenched to the extent to which it has been used in the past in hypotheses that have been successfully confirmed. Quine (1969) suggested drawing the distinction in terms of the idea of natural kinds. A completely different approach would be to point out that the reason why one thinks the observation of grue emeralds (E' or E) disconfirms the grue hypothesis (H') is because of background knowledge about constancy of color (in our usual concept of color) of many kinds of objects, and to argue that the evidence in this case should be taken as actually confirming the hypothesis H, given the Hempelian methodological fiction. It should be pointed out that the positive-instance criterion applies to a limited, though very important, kind of hypothesis and evidence: universalized conditionals for the hypothesis and positive instances for the evidence. And it supplies only a sufficient condition for confirmation. Hempel ([1945] 1965) generalized, in a natural way, this criterion to his **satisfaction criterion**, which applies to different and more complex logical structures for evidence and hypothesis and provides explicit definitions of confirmation, disconfirmation, and evidential irrelevance. Without going into any detail about this more general criterion, it is worth pointing out what Hempel took to be evidence for its adequacy. It is the satisfaction, by the satisfaction criterion, of what Hempel took to be some intuitively obvious conditions of adequacy for definitions, or *criteria*, of confirmation. Besides the

equivalence condition, two others are the entailment condition and the special-consequence condition. The entailment condition says that evidence that logically entails a hypothesis should be deemed as confirming the hypothesis. The **special-consequence condition** says that if evidence E confirms hypothesis H and if H logically entails hypothesis H', then E confirms H'. This last condition will be considered further in the section below on probabilistic approaches.

The criteria for confirmation discussed above apply in cases in which evidence reports and hypotheses are stated in the "same language," which Hempel took to be an observational language. Statements of evidence, for example, are usually referred to as observational reports in Hempel ([1945] 1965). What about confirmation of *theories*, though, which are often thought of as containing two kinds of vocabulary, observational and theoretical? **Hypothetico-deductivism (HD)** is the idea that theories and hypotheses are confirmed by their observational deductive consequences. This is different from the positive-instance criterion and the satisfaction criterion. For example, "A is an F and A is a G," which is a report of a positive instance of the hypothesis that all Fs are Gs, is not a deductive consequence of the hypothesis. The positive-instance and satisfaction criteria are formulations of the idea, roughly, that observations that are *logically consistent with* a hypothesis confirm the hypothesis, while HD says that *deductive consequences of* a hypothesis or theory confirm the hypothesis or theory.

As an example, Edmund Halley in 1705 published his prediction that a comet, now known as Halley's comet, would be visible from Earth sometime in December of 1758; he deduced this using Newtonian theory. The prediction was successful, and the December 1758 observation of the comet was taken by scientists to provide (further) very significant confirmation of Newtonian theory. Of course, the prediction was not deduced from Newtonian theory alone. In general, other needed premises include statements of *initial conditions* (in the case of the example, reports of similar or related observations at approximately 75-year intervals) and *auxiliary assumptions* (that the comet would not explode before December 1758; that other bodies in the solar system would have only an insignificant effect on the path of the comet; and so on). In addition, when the theory and the observation report share no nonlogical vocabulary (say, the theory is highly theoretical, containing no observational terms), then so-called **bridge principles** are needed to establish a deductive connection between theory and observation. An example of such a principle would be, "If there is an excess of electrons [theoretical] on the surface of a balloon [observational], then a sheet of paper [observational] will cling [observational] to it, in normal circumstances [auxiliary assumption]." Of course, if the prediction fails (an observational deductive consequence of a theory turns out to be false), then this is supposed to provide disconfirmation of the theory.

Two of the main issues or difficulties that have been discussed in connection with the HD idea have to do with what might be called distribution of credit and distribution of blame. The first has also been called the *problem of irrelevant conjunction*. If a hypothesis H logically implies an observation report E, then so does the conjunction, $H \wedge G$, where G can be any sentence whatsoever. So the basic idea of HD has the consequence that whenever an E confirms an H, the E confirms also $H \wedge G$, where G can be any (even irrelevant) hypothesis whatsoever. This problem concerns the distribution of credit. A natural response would be to refine the basic HD idea in a way to make it sensitive to the possibility that logically weaker parts of a hypothesis may suffice to deductively imply the observation report. The second issue has to do with the possibility of the failure of the prediction, of the observational deductive consequence of the hypothesis turning out to be false. This is also known as **Duhem's problem** (Duhem 1914) (see Duhem Thesis). If a hypothesis H plus statements of initial conditions I plus auxiliary assumptions A plus bridge principles B logically imply an observation report E $((H \wedge I \wedge A \wedge B) \Rightarrow E)$,and E turns out to be false, then what one can conclude is that the four-part conjunction $H \wedge I \wedge A \wedge B$ is false. And the problem is how in general to decide whether the evidence should be counted as telling against the hypothesis H, the statements of initial conditions I, the auxiliary assumptions A, or the bridge principles B.

Glymour (1980) catalogues a number of issues relevant to the assessment of HD (and accounts of confirmation in general) and proposes an alternative deductivist approach to confirmation called "the bootstrap strategy," which attempts to clarify the idea of different parts of a theory and how evidence can bear differently on them. In Glymour's bootstrap account, confirmation is a relation among a theory T, a hypothesis H, and evidence expressed as a sentence E, and Glymour gives an intricate explication of the idea that "E confirms H with respect to T," an explication that is supposed to be sensitive especially to the idea that evidence can be differently confirmationally

relevant to different hypotheses that are parts of a complex theory.

Probabilistic Approaches

One influential probabilistic approach to various issues in confirmation theory is called **Bayesian confirmation theory** (see Bayesianism). The basic idea, on which several refinements may be based, is that evidence E confirms hypothesis H if and only if the conditional probability $P(H \mid E)$ (defined as $P(H \wedge E) \mid p(E)$) is greater than the unconditional probability $P(H)$ and where $P(S)$ is the probability that the statement, or proposition, or claim, or assertion, or sentence S is true (the status of what kind of entity S might be is a concern in metaphysics or the philosophy of language, as well as in the philosophy of science). Disconfirmation is defined by reversing the inequality, and evidential irrelevance is defined by changing the inequality to an equality.

Typically in Bayesian confirmation theory, the function P is taken to be a measure of an agent's **subjective probabilities**—also called degrees of belief, partial beliefs, or degrees of confidence. Much philosophical work has been done in the area of foundations of subjective probability, the intent being to clarify or operationalize the idea that agents (e.g., scientists) have (or should have only) more or less strong or weak beliefs in propositions, rather than simply adopting the attitudes of acceptance or rejection of them. One approach, called the *Dutch book argument*, attempts to clarify the idea of subjective probability in terms of odds that one is willing to accept when betting for or against the truth or falsity of propositions (see Dutch Book Argument). Another approach, characterized as a decision theoretical approach, assumes various axioms regarding rational preference (transitivity, asymmetry, etc.) and some structural conditions (involving the richness of the set of propositions, or acts, states, and outcomes, considered by an agent) and derives, from preference data, via representation theorems, a probability assignment P and a desirability (or utility) assignment DES such that an agent prefers an item A to an item B if and only if some kind of expected utility of A is numerically greater than the expected utility of B, when the expected utilities are calculated in terms of the derived P and DES functions (see Decision Theory). Various formulas for expected utility have been proposed. (Important work in foundations of subjective probability include Ramsey 1931; de Finetti 1937; Savage [1954] 1972; Jeffrey [1965] 1983; and Joyce 1999.)

Where P measures an agent's subjective degrees of belief, $P(H|E)$ is supposed to be the agent's degree of belief in H on the assumption that E is true, or the degree of belief that the agent would have in H were the agent to learn that E is true. If a person's degree of belief in H would increase if E were learned, then it is natural to say that for this agent, E is positively evidentially relevant to H, even when E is not in fact learned. Of course, different people will have different subjective probabilities, or degrees of belief, even if the different people are equally rational, this being due to different bodies of background knowledge or beliefs possessed (albeit possibly equally justifiable or excusable, depending on one's experience), so that in this approach to confirmation theory, confirmation is a relation among three things: evidence, hypothesis, and background knowledge. The reason this approach is called Bayesian is because of the use that is sometimes made of a mathematical theorem discovered by Thomas Bayes (Bayes 1764), a simple version of which is $P(H \mid E) = P(E \mid H)P(H)/P(E)$. This is significant because it is sometimes easier to figure out the probability of an evidence statement conditional on a hypothesis than it is to figure out the probability of a hypothesis on the assumption that the evidence statement is true (for example, when the hypothesis is statistical and the evidence statement reports the outcome of an experiment to which the hypothesis applies). Bayes's theorem can be used to link these two converse conditional probabilities when the priors, $P(H)$ and $P(E)$, are known (see Bayesianism).

Bayesian confirmation theory not only provides a qualitative definition of confirmation, disconfirmation, and evidential irrelevance, but also suggests measures of degree of evidential support. The most common is the *difference measure*: $d(H,E) = P(H \mid E) - P(H)$, where confirmation, disconfirmation, and evidential irrelevance correspond to whether this measure is greater than, less than, or equal to 0, and the degree is measured by the magnitude of the difference. Another commonly used measure is the *ratio measure*: $r(H,E) = P(H \mid E)/P(H)$ where confirmation, disconfirmation, and evidential irrelevance correspond to whether this measure is greater than, less than, or equal to 1, and the degree is measured by the magnitude of the ratio. Other measures have been defined as functions of likelihoods, or converse conditional probabilities, $P(E|H)$. One application of the idea of degree of evidential support has been in the Ravens paradox, discussed above. Such definitions of degree of evidential support provide a framework within which one can clarify

intuitions that under certain conditions (contrapositive instances of the hypothesis that all ravens are black [i.e., nonblack nonravens]) confirm the hypothesis that all ravens are black, but to a minuscule degree compared with positive instances (black ravens).

What follows is a little more detail about the application of Bayesian confirmation theory to the positive-instance criterion and the Ravens paradox. (see Eells 1982 for a discussion and references.) Let H be the hypothesis that all ravens are black; let R_A symbolize the statement that object A is a raven; and let B_A symbolize the statement that object A is black. It can be shown that if H is probabilistically independent of R_A, (i.e., $P(H|R_A) = P(H)$), then a positive instance (or report of one), $R_A \wedge B_A$, of H confirms H in the Bayesian sense (i.e., $P(H|R_A \wedge B_A) > P(H)$) if and only if $P(B_A|R_A) < 1$ (which latter inequality can naturally be interpreted as saying that it was not *already* certain that A would be black if a raven). Further, on the same independence assumption, it can be shown that a positive instance, $R_A \wedge B_A$, confirms H more than a contrapositive instance, $\neg B_A \wedge \neg R_A$ if and only if $P(B_A|R_A) < P(\neg R_A|\neg B_A)$.

What about the assumption of probabilistic independence of H from R_A? I. J. Good (1967) has proposed counterexamples to the positive-instance criterion like the following. Suppose it is believed that *either* (1) there are just a few ravens in the world and they are all black *or* (2) there are lots and lots of ravens in the world, a very, very few of which are nonblack. Observation of a raven, even a black one (hence a positive instance of H), would tend to support supposition 2 against supposition 1 and thus undermine H, so that a positive instance would disconfirm H. But in this case the independence assumption does not hold, so that Bayesian confirmation theory can help to isolate the kinds of situations in which the positive-instance criterion holds and the kinds in which it may not.

Bayesian confirmation theory can also be used to assess Hempel's proposed conditions of adequacy for criteria of confirmation. Recall, for example, his special-consequence condition, discussed above: If E confirms H and H logically entails H', then E must also confirm H'. It is a theorem of probability theory that if H logically entails H', then $P(H')$ is at least as great as $P(H)$. If the inequality is strict, then an E can increase the probability of H while decreasing the probability of H', consistent with H logically entailing H'. This fact can be used to construct intuitively compelling examples of an H entailing an H' and an E confirming the H while disconfirming the H', telling against the special-consequence condition and also in favor of Bayesian confirmation theory (e.g., Eells 1982).

Some standard objections to Bayesian confirmation theory are characterized as the *problem of the priors* and the *problem of old evidence*. As to the first, while it is sometimes admitted that it makes sense to assign probabilities to evidence statements E conditional on some hypothesis H (even in the absence of much background knowledge), it is objected that it often does not make sense to assign unconditional, or "prior," probabilities to hypothesis H or to evidence statements (reports of observation) E. If H is a newly formulated physical hypothesis, for example, it is hard to imagine what would *justify* an assignment of probability to it prior to evidence—but that is just what the suggested criterion of confirmation, Bayes's theorem, and the measures of confirmation described above require. Such issues make some favor a **likelihood** approach to the evaluation of evidence—Edwards (1972) and Royall (1997), for instance, who represent a different approach and tradition in the area of statistical inference. According to one formulation of the likelihood account, an E confirms an H more than the E confirms an H' if and only if $P(E|H)$ is greater than $P(E|H')$. This is a comparative principle, an approach that separates the question of which hypothesis it is more justified to believe given the evidence (or the comparative acceptability of hypotheses given the evidence) from the question of what the comparative significance is of evidence for one hypothesis compared with the evidence's significance for another hypothesis. It is the latter question that the likelihood approach actually addresses, and it is sometimes suggested that the degree to which an E supports an H compared with the support of E for an H' is measured by the likelihood ratio, $P(E|H)/P(E|H')$. Also, likelihood measures of *degree* of confirmation of *single* hypotheses have been proposed, such as $L(H, E) = P(E|H)/P(E|\neg H)$, or the log of this ratio. (see Fitelson 2001 and Forster and Sober 2002 for recent discussion and references.)

Another possible response to the problem of priors is to point to convergence theorems (as in de Finetti 1937) and argue that initial settings of priors does not matter in the long run. According to such theorems, if a number of agents set different priors, are exposed to the same series of evidence, and update their subjective probabilities (degrees of belief) in certain ways, then, almost certainly, their subjective probabilities will

eventually converge on each other and, under certain circumstances, upon the truth.

The problem of old evidence (Glymour 1980; Good 1968, 1985) arises in cases in which $P(E) = 1$. It is a theorem of probability theory that in such cases $P(H|E) = P(H)$, for any hypothesis H, so that in the Bayesian conception of confirmation as formulated above, an E with probability 1 cannot confirm any hypothesis H. But this seems to run against intuition in some cases. An often-cited such case is the confirmation that Einstein's general theory of relativity apparently was informed by already known facts about the behavior of the perihelion of the planet Mercury. One possible Bayesian solution to the problem, suggested by Glymour (1980), would be to say that it is not the already known E that confirms the H after all, but rather a newly discovered logical or explanatory relation between the H and the E. Other solutions have been proposed, various versions of the problem have been distinguished (see Earman 1992 for a discussion), and the problem remains one of lively debate.

ELLERY EELLS

References

Bayes, Thomas (1764), "An Essay Towards Solving a Problem in the Doctrine of Chance," *Philosophical Transactions of the Royal Society of London* 53: 370–418.

de Finetti, Bruno (1937), "La prévision: Ses lois logiques, ses sources subjectives," *Annales de l'Institute Henri Poincaré* 7: 1–68.

Duhem, Pierre (1914), *The Aim and Structure of Physical Theory*. Princeton, NJ: Princeton University Press.

Earman, John (1992), *Bayes or Bust: A Critical Examination of Bayesian Confirmation Theory*. Cambridge, MA, and London: MIT Press.

Edwards, A. W. F. (1972), *Likelihood*. Cambridge: Cambridge University Press.

Eells, Ellery (1982), *Rational Decision and Causality*. Cambridge and New York: Cambridge University Press.

Fitelson, Branden (2001), *Studies in Bayesian Confirmation Theory*. Ph.D. dissertation, University of Wisconsin–Madison.

Forster, Malcolm, and Elliott Sober (2002), "Why Likelihood?" in M. Taper and S. Lee (eds.), *The Nature of Scientific Evidence*. Chicago: University of Chicago Press.

Glymour, Clark (1980), *Theory and Evidence*. Princeton, NJ: Princeton University Press.

Good, I. J. (1967), "The White Shoe Is a Red Herring," *The British Journal for the Philosophy of Science* 17: 322.

——— (1968), "Corroboration, Explanation, Evolving Probability, Simplicity, and a Sharpened Razor," *British Journal for the Philosophy of Science* 19: 123–143.

——— (1985), "A Historical Comment Concerning Novel Confirmation," *British Journal for the Philosophy of Science* 36: 184–186.

Goodman, Nelson (1965), *Fact, Fiction, and Forecast*. New York: Bobbs-Merrill.

Hempel, Carl G. ([1945] 1965), "Studies in the Logic of Confirmation," in *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press, 3–51. Originally published in *Mind* 54: 1–26 and 97–121.

Jeffrey, Richard C. ([1965] 1983), *The Logic of Decision*. Chicago and London: University of Chicago Press.

Joyce, James M. (1999), *The Foundations of Causal Decision Theory*. Cambridge and New York: Cambridge University Press.

Nicod, Jean (1930), *Foundations of Geometry and Induction*. New York and London: P. P. Wiener.

Quine, W. V. (1969), "Natural Kinds," in *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Ramsey, Frank Plumpton (1931), "Truth and Probability," in R. B. Braithwaite (ed.), *The Foundations of Mathematics and Other Logical Essays*. London: Routledge and Kegan Paul, 156–198.

Royall, Richard (1997), *Statistical Evidence: A Likelihood Paradigm*. Boca Raton, FL: Chapman and Hall.

Savage, Leonard ([1954] 1972), *The Foundations of Statistics*. New York: Dover Publications.

See also **Bayesianism; Carnap, Rudolf; Decision Theory; Dutch Book Argument; Epistemology; Hempel, Carl Gustav; Induction, Problem of; Inductive Logic; Probability**

CONNECTIONISM

Connectionist models, also known as models of **parallel distributed processing** (PDP) and **artificial neural networks** (ANN), have merged into the mainstream of cognitive science since the mid-1980s. Connectionism currently represents one of two dominant approaches (symbolic modeling is the other) within artificial intelligence used to develop computational models of mental processes.

Unlike symbolic modeling, connectionist modeling also figures prominently in computational neuroscience (where the preferred term is *neural network modeling*).

Connectionist modeling first emerged in the period 1940–1960, as researchers explored possible approaches to using networks of simple neurons to perform psychological tasks, but it fell into decline when limitations of early network designs became apparent around 1970. With the publication of the PDP Research Group volumes (Rumelhart, McClelland, et al. 1986; McClelland, Rumelhart, et al. 1986), connectionism was rescued from over a decade of popular neglect, and the way was opened for a new generation to extend the approach to fresh explanatory domains. (For a collection of historically significant papers from these neglected years and before, see Anderson and Rosenfeld 1988; Anderson, Pellionisz, and Rosenfeld 1990.) Despite some early claims that connectionism constitued a new, perhaps revolutionary, way of understanding cognition, the veritable flood of network-based research has ultimately occurred side by side with other, more traditional, styles of modeling and theoretical frameworks.

The renaissance in connectionist modeling is a result of convergence from many different fields. Mathematicians and computer scientists attempt to describe the formal, mathematical properties of abstract network architectures. Psychologists and neuroscientists use networks to model behavioral, cognitive, and biological phenomena. Roboticists also make networks the control systems for many kinds of embodied artificial agents. Finally, engineers employ connectionist systems in many industrial and commercial applications. Research has thus been driven by a broad spectrum of concerns, ranging from the purely theoretical to problem-solving applications for problems in various scientific domains to application-based or engineering needs.

Given these heterogeneous motivations, and the recent proliferation of network models, analytic techniques, applications, and theories, it is appropriate to ask whether connectionism constitutes a coherent research program or is instead primarily a modeling tool. Following Lakatos, connectionism could be construed as a research program involving a set of core theoretical principles about the role of networks in explaining and understanding cognition, a set of positive and negative heuristics that guide research, an ordering of the important commitments of connectionist modeling, and a set of principles and strategies dictating how recalcitrant empirical results are to be accounted for (see Lakatos, Imre; Research Programs). The greater the disunity in these factors, the less connectionism resembles a research program, and the more it appears to be a convenient tool for modeling certain phenomena. If it is a modeling tool, connectionism need not commit modelers to having anything in common beyond their use of the particular mathematical and formal apparatus itself.

This article will briefly describe the features of prevalent connectionist architectures and discuss a number of challenges to the use of these models. One challenge comes from symbolic models of cognition, which present an alternative representational framework and set of processing assumptions. Another comes from a purportedly nonrepresentational framework, that of nonlinear dynamical systems theory. Finally, there is the neuroscientific challenge to the disciplinary boundaries drawn around "cognitive" modeling by some connectionist psychologists. The status of connectionism is assessed in light of these challenges.

The Properties of Connectionist Models

Connectionist networks are built up from basic computational elements called units or nodes, which are linked to each other via weighted connections, called simply **weights**. Units take on a variable numerical level of activation, and they pass activation to each other via the weights. Weights determine how great an effect one unit has on other units. This effect may be positive (excitatory) or negative (inhibitory). The net input a unit receives at a time is the weighted sum of the activations on all of the units that are active and connected to it. Given the net input, an activation function (often nonlinear or imposing a threshold) determines the activation of the unit. In this way, activation is passed in parallel throughout the network. Connectionist networks compute functions by mapping vectors of activation values onto other such vectors.

Multilayer **feedforward networks** are the most intensively studied and widely used class of contemporary models. Units are arranged into layers, beginning with an input layer and passing through a number of intermediate hidden layers, terminating with an output layer. There are no reciprocal connections, so activation flows unidirectionally through the network. The modeler assigns representational significance to the activation vectors at the input and output layers of a network, thereby forging the link between the model and the cognitive task to be explained.

For example, Sejnowski and Rosenberg's NETtalk architecture is designed to map graphemic inputs (letters) onto phonemic outputs. It consists of an input layer of seven groups of 29 units, a hidden layer of 80 units, and an output layer of 26 units. Each layer is completely connected to the next one. Vectors of activity in the network represent aspects of the letter-reading task. The input groups are used to represent the seven letters of text that the network is perceiving at a time, and the output layer represents the phoneme corresponding to the fourth letter in the input string. The task of the network is to pronounce the text correctly, given the relevant context. When the values of the weights are set correctly, the network can produce the appropriate phoneme-representing vectors in response to the text-representing vectors.

Simple networks can be wired by hand to compute some functions, but in networks containing hundreds of units, this is impossible. Connectionist systems are therefore usually not programmed in the traditional sense, but are trained by fixing a learning rule and repeatedly exposing the network to a subset of the input-output mappings it is intended to learn. The rule then systematically adjusts the network's weights until its outputs are near the target output values. One way to classify learning rules is according to whether they require an external trainer. Unsupervised learning (e.g., Hebbian learning) does not require an external trainer or source of error signals. Supervised learning, on the other hand, requires that something outside the network indicate when its performance is incorrect. The most popular supervised learning rule currently in use is the **backpropagation** rule.

In backpropagation learning, the network's weights are initially set to random values (within certain bounds). The network is then presented with patterns from the training environment. Given random weights, the network's response will likely be far from the intended mapping. The difference between the output and the target is computed by the external trainer and used to send an error signal backward through the network. As the signal propagates, the weights between each layer are adjusted by a slight amount. Over many training cycles, the network's performance gradually approaches the target. When the output is within some criterion distance of the target, training ceases. Since the error is being reduced gradually, backpropagation is an instance of a gradient-descent learning algorithm.

Backpropagation-trained networks have been successful at performing in many domains, including past-tense transformation of verbs, generation of prototypes from exemplars, single word reading, shape-from-shading extraction, visual object recognition, modeling deficits arising in deep dyslexia, and more. Their formal properties are well known. However, they suffer from a number of problems. Among these is the fact that learning via backpropagation is extremely slow, and increasing the learning-rate parameter typically results in overshooting the optimum weights for solving the task.

Another problem facing feedforward networks generally is that individual episodes of processing inputs are independent of each other except for changes in weights resulting from learning. But often a cognitive agent is sensitive not just to what it has learned over many episodes, but to what it processed recently (e.g., the words prior to the preceding one). The primary way sensitivity to context has been achieved in feedforward networks has been to present a constantly moving window of input. For example, in NETtalk, the input specified the three phonemes before and three phonemes after the one to be pronounced. But this solution is clearly a kludge and suffers from the fact that it imposes a fixed window. If sensitivity to the item four back is critical to correct performance, the network cannot perform correctly.

An alternative architecture that is increasingly being explored is the **simple recurrent network** (SRN) (Elman 1991). SRNs have both feedforward and recurrent connections. In the standard model, an input layer sends activity to a hidden layer, which has two sets of outgoing connections: one to other hidden layers and eventually on to the output layer, and another to a specialized context layer. The weights to the units in the context layer enable it to construct a copy of the activity in the hidden units. The activation over these units is then treated as an additional input to the same hidden units at the next temporal stage of processing. This allows for a limited form of short-term memory, since activity patterns that were present during the previous processing cycle have an effect on the next cycle. Since the activity on the previous cycle was itself influenced by that on a yet earlier cycle, this allows for memory extending over several previous processing epochs (although sensitivity to more than one cycle back will be diminished).

Once trained in a variation of backpropagation, many SRNs are able to discover patterns in temporally ordered sets of events. Elman (1991) has trained SRNs on serially presented words in an attempt to teach them to predict the grammatical category of the next word in a sentence. The networks can achieve fairly good performance at

this task. Since the networks were never supplied information about grammatical categories, this suggests that they induced a representation of a more abstract similarity among words than was present in the raw training data. There are many other kinds of neural network architecture. (For further details on their properties and applications, see Anderson 1995; Bechtel and Abrahamsen 2002.)

Connectionism and Symbolic Models

Within cognitive science, symbolic models of cognition have constituted the traditional alternative to connectionism. In symbolic models, the basic representational units are symbols having both syntactic form and typically an intuitive semantics that corresponds to the elements picked out by words of natural language. The symbols are discrete and capable of combining to form complex symbols that have internal syntactic structure. Like the symbol strings used in formal logic, these complex symbols exhibit variable binding and scope, function-and-argument structure, cross-reference, and so on. The semantics for complex symbols is combinatorial: The meaning of a complex symbol is determined by the meanings of its parts plus its syntax. Finally, in symbolic models the dynamics of the system are governed by rules that transform symbols into other symbols by responding to their syntactic or formal properties. These rules are intended to preserve the truth of the structures manipulated. Symbolic models are essentially proof-theoretic engines.

Connectionist models typically contain units that do not individually represent lexicalized semantic contents. (What are called *localist* networks are an exception. In these, individual units are interpretable as expressing everyday properties or propositions. See Page 2000.) More commonly, representations with lexicalized content are *distributed* over a number of units in a network (Smolensky 1988). In a distributed scheme, individual units may stand for repeatable but nonlexicalized microfeatures of familiar objects, which are themselves represented by vectors of such features. In networks of significant complexity, it may be difficult, if not impossible, to discern what content a particular unit is carrying.

In networks, there is no clear analog to the symbolicist's syntactic structures. Units acquire and transmit activation values, resulting in larger patterns of coactivation, but these patterns of units do not themselves syntactically compose. Also, there is no clear program/data distinction in connectionist systems. Whether a network is hand-wired or trained using a learning rule, the modifications are changes to the weights between units. The new weight settings determine the future course of activation in the network and simultaneously constitute the data stored in the network. There are no explicitly represented rules that govern the system's dynamics.

Classical theorists (Fodor and Pylyshyn 1988) have claimed that there are properties of cognition that are captured naturally in symbolic models but that connectionist models can capture them only in an ad hoc manner, if at all. Among these properties are the productivity and the systematicity of thought. Like natural language, thought is productive, in that a person can think a potentially infinite number of thoughts. For example, one can think that Walt is an idiot, that Sandra believes that Walt is an idiot, that Max wonders whether Sandra believes that Walt is an idiot, and so on. Thought is also systematic, in that a person who can entertain a thought can also entertain many other thoughts that are semantically related to it (Cummins 1996). If a person can think that Rex admires the butler's courage, that person can also think that the butler admires Rex's courage. Anyone who can think that dogs fear cats can think that cats fear dogs, and so on. Unless each thought is to be learned anew, these capacities need some finite basis.

Symbolicists argue that this basis is compositionality: Thought possesses a combinatorial syntax and semantics, according to which complex thoughts are built up from their constituent concepts, and those concepts completely determine the meaning of a complex thought. The compositionality of thought would explain both productivity and systematicity. Grasping the meaning of a set of primitive concepts and grasping the recursive rules by which they can be combined is sufficient for grasping the infinite number of thoughts that the concepts and rules can generate. Similarly, grasping a syntactic schema and a set of constituent concepts explains why the ability to entertain one thought necessitates the ability to entertain others: Concepts may be substituted into any permissible role slot in the schema.

The challenge symbolicists have put to connectionists is to explain productivity and systematicity in a principled fashion, without merely implementing a symbolic architecture on top of the connectionist network (for one such implementation, see Franklin and Garzon 1990). One option that some connectionists pursue is to deny that thought is productive or systematic, as symbolicists

characterize these properties. For example, one might deny that it is possible to think any systematically structured proposition. Although one can compose the symbol string ''The blackberry ate the bear,'' that does not entail that one is able to think such a thought. But clearly much of thought exhibits some degree of productivity and systematicity, and it is incumbent on connectionists to offer some account of how it is achieved.

Connectionists have advanced a number of proposals for explaining productivity and systematicity. Two of the most widely discussed are Pollack's **recursive auto-associative memories** (RAAMs) and Smolensky's **tensor product networks** (Pollack 1990; Smolensky, 1991). RAAMs are easier to understand. An auto-associative network is one trained to produce the same pattern on the output layer as is present on the input layer. If the hidden layer is smaller than the input and output layers, then the pattern produced on the hidden layer is a compressed representation of the input pattern. If the input layer contains three times the number of units as the hidden layer, then one can treat the input activation as consisting of three parts constituting three patterns (e.g., for different words) and recursively compose the hidden pattern produced by three patterns with two new ones. In such an encoding, one is implicitly ignoring the output units, but one can also ignore the input units and supply patterns to the hidden units, allowing the RAAM to generate a pattern on the output units. What is interesting is that even after several cycles of compression, one can, by recursively copying the pattern on one-third of the output units back onto the hidden units, recreate with a fair degree of accuracy the original input patterns.

If RAAMs are required to perform many cycles of recursive encoding, the regeneration of the original pattern may exhibit errors. But up to this point, the RAAM has exhibited a degree of productivity by composing representations of complex structures from representations of their parts. One can also use the compressed patterns in other processing (e.g., to train another feedforward network to construct the compressed representation of a passive sentence from the corresponding active sentence). RAAMs thus exhibit a degree of systematicity. But the compressed representations are not composed according to syntactic principles. Van Gelder (1990), accordingly, construes them as manifesting functional, not explicit, compositionality.

Symbolicists have rejected such functional compositionality as inadequate for explaining cognition. In many respects, this debate has reached a standoff. In part its resolution will turn on the issue posed earlier: To what degree do humans exhibit productivity and systematicity? But independently of that issue, there are serious problems in scaling up from connectionist networks designed to handle toy problems to ones capable of handling the sort of problems humans deal with regularly (e.g., communicating in natural languages). Thus, it is not clear whether solutions similar to those employed in RAAMs (or in SRNs, which also exhibit a degree of productivity and systematicity) will account for human performance. (Symbolic models have their own problems in scaling, and so are not significantly better off in practice.)

Connectionism and Dynamical Systems Theory

Although the conflict between connectionists and symbolicists reached a stalemate in the 1990s, within the broader cognitive science community a kind of accord was achieved. Connectionist approaches were added to symbolic approaches as parts of the modeling toolkit. For some tasks, connectionist models proved to be more useful tools than symbolic models, while for others symbolic models continued to be preferred. For yet other tasks, connectionist models were integrated with symbolic models into hybrids.

As this was happening, a new competitor emerged on the scene, an approach to cognition that challenged both symbolic and connectionist modeling insofar as both took seriously that cognitive activity involved the use of some form of representation. **Dynamical systems theory** suggested that rather than construing cognition as involving syntactic manipulation of representations or processing them through layers of a network, one should reject the notion of representation altogether. The alternative these critics advanced was to characterize cognitive activity in terms of a (typically small) set of variables and to formulate (typically nonlinear) equations that would relate the values of different variables in terms of how they changed over time. In physics, dynamical systems theory provides a set of tools for understanding the changes over time of such systems of variables. For example, each variable can be construed as defining a dimension in a multidimensional *state space*, and many systems, although starting at different points in this space, will settle onto a fixed point or into a cycle of points. These points are known as *attractors*, and the paths to them as the *transients*. The structure of the state space can often be represented

geometrically—for example, by showing different basins, each of which represents starting states that will end up in a different attractor.

Connectionist networks, especially those employing recurrent connections, are dynamical systems, and some connectionist modelers have embraced the tools of dynamical systems theory for describing their networks. Elman, for example, employs such tools to understand how networks manage to learn to respect syntactic categories in processing streams of words. Others have made use of some of the more exotic elements in dynamical systems theory, such as activitation functions that exhibit deterministic chaos, to develop new classes of networks that exhibit more complex and interesting patterns of behavior than simpler networks (e.g., jumping intermittently between two competing interpretations of an ambiguous perceptual figure such as the duck-rabbit (see van Leeuwen, Verver, and Brinkers 2000)). Some particularly extreme dynamicists, however, contend that once one has characterized a cognitive system in this way, there is no further point to identifying internal states as representations and characterizing changes as operations on these representations. Accordingly, they construe dynamical systems theory as a truly radical paradigm shift for cognitive science (van Gelder 1995).

This challenge has been bolstered by some notable empirical successes in dynamical systems modeling of complex cognition and behavior. For example, Busemeyer and Townsend (1993) offer a model of decision under uncertainty, decision field theory, that captures much of the empirical data about the temporal course of decision making using difference equations containing only seven parameters for psychological quantities such as attention weight, valence, preference, and so on. Connectionist models, by contrast, have activation state spaces of as many dimensions as they have units. There are dynamical systems models of many other phenomena, including coordination of finger movement, olfactory perception, infants' stepping behavior, the control of autonomous robot agents, and simple language processing. The relative simplicity and comprehensibility of their models motivates the antirepresentational claims advanced by dynamical systems theorists.

There is reason, though, to question dynamical systems theory's more radical challenge to cognitive modeling. One can accept the utility of characterizing a system in terms of a set of equations and portraying its transitions through a multidimensional state space without rejecting the utility of construing states within the system as representational and the trajectories through state space in terms of transitions between representations. This is particularly true when the system is carrying out what one ordinarily thinks of as a complex reasoning task such as playing chess, where the task is defined in terms of goals and the cognizer can be construed as considering different possible moves and their consequences. To understand why a certain system is able to play chess successfully, rather than just recognizing that it does, the common strategy is to treat some of its internal states as representing goals or possible moves. Moreover, insofar as these internal states are causally connected in appropriate ways, they do in fact carry information (in what is fundamentally an informational-theoretic sense, in which the state covaries with referents external to the system), which is then utilized by other parts of the system. In systems designed to have them, these information-carrying states may arise without their normal cause (as when a frog is subjected to a laboratory with bullets on a string moving in front of its eyes). In these situations the system responds as it would if the state were generated by the cause for which the response was designed or selected. Such internal states satisfy a common understanding of what a representation is, and the ability to understand why the system works successfully appeals to these representations. If this construal is correct, then dynamical systems theory as well is not an alternative to connectionist modeling, but should be construed as an extension of the connectionist toolkit (Bechtel and Abrahamsen, 2002).

Connectionism and Neuroscience

Connectionist models are often described as being *neurally inspired*, as the term "artificial neural networks" implies. More strongly, many connectionists have claimed that their models enjoy a special sort of *neural plausibility*. If there is a significant similarity between the processing in ANNs and the activity in real networks of neurons, this might support connectionism for two reasons. First, since the cognitive description of the system closely resembles the neurobiological description, it seems that connectionist models in psychology have a more obvious account about how the mind might supervene on the brain than do symbolic models (see Supervenience). Second, connectionist models might provide a fairly direct characterization of the functioning of the neural level itself (e.g., the particular causal interactions among neurons). This functioning is not easily revealed by many standard neuroscience methods. For example, localization of mental activity via functional

magnetic resonance imaging can indicate which brain regions are preferentially activated during certain tasks, but this alone does not give information about the specific computations being carried out within those regions. Connectionist modeling of brain function thus might supplement other neurobiological techniques.

This strategy can be illustrated within the domain of learning and memory. Neuropsychological studies suggest that the destruction of structures in the medial temporal lobes of the brain, especially the hippocampus, results in a characteristic pattern of memory deficits. These include (i) profound anterograde amnesia for information presented in declarative or verbal form, such as arbitrary paired associates, as well as memory for particular experienced events more generally (episodic memory); (ii) preserved implicit memory for new information, such as gradually acquired perceptual-motor skills; and (iii) retrograde amnesia for recently acquired information, with relative preservation of memories farther back in time. This triad of deficits was famously manifested by H.M., a patient who underwent bilateral removal of sections of his medial temporal lobes in the early 1950s in order to cure intractable epilepsy. Since H.M., this pattern of deficits has been confirmed in other human and animal studies.

McClelland, McNaughton, and O'Reilly (1995) offer an explanation of these deficits based on connectionist principles. One feature of backpropagation-trained networks is that once they are trained on one mapping, they cannot learn another mapping without "unlearning" the previously stored knowledge. This phenomenon is known as **catastrophic interference**. However, catastrophic interfence can be overcome if, rather than fully training a network on one mapping, then training it on another, the training sets are interleaved so that the network is exposed to both mappings in alternation. When the training environment is manipulated in this way, the network can learn both mappings without overwriting either one.

As a model of all learning and memory, this technique suffers from being slow and reliant on a fortunate arrangement of environmental contingencies. However, McClelland et al. (1995) suggest that learning in the neocortex may be characterized by just such a process, if there is a neural mechanism that stores, organizes, and presents appropriately interleaved stimuli to it. They conjecture that this is the computational function of the hippocampus. Anatomically, the hippocampus receives convergent inputs from many sensory centers and has wide-ranging efferent connections to the neocortex. The pattern of deficits resulting from hippocampal lesions could be explained on the assumption that the hippocampus has a method of temporarily storing associations among stimuli without catastrophic interference. Ablation of the hippocampus results in anterograde amnesia for arbitrary associations and declarative information because the neocortex alone is incapable of learning these without the appropriately interleaved presentation. Implicit learning is preserved because it does not require rapid integration of many disparate representations into a single remembered experience; further, it typically takes many trials for mastery, as is also the case with backpropagation learning. Finally, the temporally graded retrograde amnesia is explained by the elimination of memory traces that are temporarily stored in the hippocampus itself. Older memories have already been integrated into the neocortex, and hence are preserved.

McClelland et al. (1995) did not model the hippocampus directly; rather, they implemented it as a black box that trained the neocortical network according to the interleaving regimen. Others have since presented more elaborate models. Murre, Graham, and Hodges (2001) have implemented a system called TraceLink that features a network corresponding to a simplified single-layer hippocampus, the neocortex, and a network of neuromodulatory systems (intended to correspond to the basal forebrain nuclei). TraceLink accounts for the data reviewed by McClelland et al. (1995), and also predicts several phenomena associated with semantic dementia. Other models incorporating a more elaborate multilayer hippocampal network have been presented by Rolls and Treves (1998) and O'Reilly and Rudy (2001). These models collectively support the general framework set out by McClelland et al. (1995) concerning the computational division of labor between the hippocampus and the neocortex in learning and memory.

These studies of complementary learning systems suggest a useful role for network-based modeling in neuroscience. However, this role is presently limited in several crucial respects. The models currently being offered are highly impoverished compared with the actual complexity of the relevant neurobiological structures. Assuming that these networks are intended to capture neuron-level interactions, they are several orders of magnitude short of the number of neurons and connections in the brain. Further, the backpropagation rule itself is biologically implausible if interpreted at the neuronal level, since it allows individual weights to take on either positive or negative values, while actual axonal connections are either excitatory or

inhibitory, but never both. There is no network model that captures all of the known causal properties of neurons, even when the presence of glial cells, endocrine regulators of neural function, and other factors are abstracted away.

A common response to these objections is to interpret networks as describing only select patterns of causal activity among large populations of neurons. In many cases, this interpretation is appropriate. However, there are many kinds of network models in neuroscience, and they can be interpreted as applying to many different levels of organization within the nervous system, including individual synaptic junctions on dendrites, particular neurons within cortical columns, and interactions at the level of whole neural systems. No single interpretation appears to have any special priority over the others. The specific details of the network architecture are dictated in most cases by the particular level of neural analysis being pursued, and theorists investigate multiple levels simultaneously. There are likely to be at least as many distinct kinds of possible connectionist models in neuroscience as there are distinct levels of generalization within the nervous system.

Conclusions

At the beginning of this article, the question was asked whether connectionism is best thought of as a research program, a modeling tool, or something in between. Considering the many uses to which networks have been put, and the many disciplines that have been involved in cataloguing their properties, it seems unlikely that there will be any common unity of methods, heuristics, principles, etc., among them. Ask whether a neuroscientist using networks to model the development of receptive fields in the somatosensory system would have anything in common with a programmer training a network to take customers' airline reservations. Each of these might be a neural network theorist, despite having nothing significant in common besides the formal apparatus they employ. Across disciplines, then, connectionism lacks the characteristic unity one would expect from a research program.

The question may be asked again at the level of each individual discipline. This article has not surveyed every field in which networks have played a significant role but has focused on their uses in artificial intelligence, psychology, and neuroscience. Even within these fields, the characteristic unity of a research program is also largely absent, if one takes into account the diverse uses that are made of networks.

In psychology, for instance, there are some connectionists who conceive of their models as providing a theory of how mental structures might functionally resemble, and therefore plausibly supervene in, the organization of large-scale neuronal structures (see Psychology). However, there are just as many theorists who see their work as being only, in some quite loose sense, neurally inspired. The organization of the NETtalk network is not particularly neurally plausible, since it posits a simple three-layered linear causal process leading from the perception of letters to the utterance of phonemes. Being a connectionist in psychology does not appear to require agreement on the purpose of the models used or the possible data (e.g., neurobiological) that might confirm or disconfirm them. This is what one might expect of a tool rather than a research program.

It is perhaps ironic that this state of affairs was predicted by Rumelhart, one of the theorists who revitalized connectionism during the 1980s. In a 1993 interview, Rumelhart claimed that as networks become more widely used in a number of disciplines, "there will be less and less of a core remaining for neural networks per se and more of, 'Here's a person doing good work in [her] field, and [she's] using neural networks as a tool'" (Anderson and Rosenfeld 1998, 290). Within the fields, in turn, network modeling will "[d]isappear as an identifiable separate thing" and become "part of doing science or doing engineering" (291). Such a disappearance, however, may not be harmful. Connectionist networks, like other tools for scientific inquiry, are to be evaluated by the quality of the results they produce. In this respect, they have clearly proven themselves a worthy, and sometimes indispensible, component of research in an impressive variety of disciplines.

Dan Wieskopf and William Bechtel

References

Anderson, James A. (1995), *An Introduction to Neural Networks*. Cambridge, MA: MIT Press.

Anderson, James A., and Edward Rosenfeld (eds.) (1988), *Neurocomputing: Foundations of Research*, vol. 1. Cambridge, MA: MIT Press.

——— (1998), *Talking Nets: An Oral History of Neural Networks*. Cambridge, MA: MIT Press.

Anderson, James A., Andras Pellionisz, and Edward Rosenfeld (eds.) (1990), *Neurocomputing: Directions for Research*, vol. 2. Cambridge, MA: MIT Press.

Bechtel, William, and Adele Abrahamsen (2002), *Connectionism and the Mind: Parallel Processing, Dynamics, and Evolution in Networks*. Oxford: Basil Blackwell.

Busemeyer, Jerome R., and James T. Townsend (1993), "Decision Field Theory: A Dynamic-Cognitive

Approach to Decision Making in an Uncertain Environment,'' *Psychological Review* 100: 423–459.
Cummins, Robert (1996), ''Systematicity,'' *Journal of Philosophy* 93: 591–614.
Elman, Jeffrey L. (1991), ''Distributed Representations, Simple Recurrent Networks, and Grammatical Structure,'' *Machine Learning* 7: 195–225.
Fodor, Jerry A., and Zenon Pylyshyn (1988), ''Connectionism and Cognitive Architecture: A Critical Analysis,'' *Cognition* 28: 3–71.
Franklin, Stan, and Max Garzon (1990), ''Neural Computability,'' in O. M. Omidvar (ed.), *Progress in Neural Networks*, vol. 1. Norwood, NJ: Ablex, 127–145.
McClelland, James L., Bruce L. McNaughton, and Randall C. O'Reilly (1995), ''Why There Are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory,'' *Psychological Review* 102: 419–457.
McClelland, James L., David E. Rumelhart, and the PDP Research Group (1986), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 2. Cambridge, MA: MIT Press.
Murre, Jacob, Kim S. Graham, and John R. Hodges (2001), ''Semantic Dementia: Relevance to Connectionist Models of Long-Term Memory,'' *Brain* 124: 647–675.
O'Reilly, Randall C., and J. W. Rudy (2001), ''Conjunctive Representations in Learning and Memory: Principles of Cortical and Hippocampal Function,'' *Psychological Review* 108: 311–345.
Page, M. (2000), ''Connectionist Modeling in Psychology: A Localist Manifesto,'' *Behavioral and Brain Sciences* 23: 443–512.
Pollack, Jordan B (1990), ''Recursive Distributed Representations,'' *Artificial Intelligence* 46: 77–105.
Rolls, Edmund T., and Alessandro Treves (1998), *Neural Networks and Brain Function*. Oxford: Oxford University Press.
Rumelhart, David E., James L. McClelland, and the PDP Research Group (1986), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1. Cambridge, MA: MIT Press.
Smolensky, Paul (1988), ''On the Proper Treatment of Connectionism,'' *Behavioral and Brain Sciences* 11: 1–74.
——— (1991), ''Connectionism, Constituency, and the Language of Thought,'' in Barry Loewer and Georges Rey (eds.), *Meaning in Mind: Fodor and His Critics*. Oxford: Basil Blackwell.
van Gelder, Timothy (1990), ''Compositionality: A Connectionist Variation on a Classical Theme,'' *Cognitive Science* 14: 355–384.
——— (1995), ''What Might Cognition Be, If Not Computation?'' *Journal of Philosophy* 111: 345–381.
van Leeuwen, Cees, S. Verver, and M. Brinkers (2000), ''Visual Illusions, Solid/Outline Invariance and Non-Stationary Activity Patterns,'' *Connection Science* 12: 279–297.

See also **Artificial Intelligence; Cognitive Science; Neurobiology; Psychology, Philosophy of; Supervenience**

CONSCIOUSNESS

Consciousness is extremely familiar, yet it is at the limits—beyond the limits, some would say—of what one can sensibly talk about or explain. Perhaps this is the reason its study has drawn contributions from many fields, including psychology, neuroscience, philosophy, anthropology, cultural and literary theory, artificial intelligence, physics, and others. The focus of this article is on the varieties of consciousness, different problems that have been raised about these varieties, and prospects for progress on these problems.

Varieties of Consciousness

Creature vs. State Consciousness

One attributes consciousness both to people and to their psychological states. An agent can be conscious (as opposed to unconscious), and that agent's desire for a certain emotional satisfaction might be unconscious (as opposed to conscious). Rosenthal (1992) calls the former **creature consciousness** and the latter **state consciousness**. Most, but not all, discussion of consciousness in the contemporary literature concerns state consciousness rather than creature consciousness. Rosenthal (1992) goes on to propose an explanation of state consciousness in terms of creature consciousness, according to which a state is conscious just in case an agent who is in the state is conscious of it—but this proposal has proved controversial (e.g., Dretske 1993).

Essential vs. Nonessential Consciousness

Focusing on conscious states, one may distinguish those that are essentially conscious from

those that are (or might be) conscious but not essentially so. The distinction is no doubt vague, but, to a first approximation, a state is essentially conscious just in case being in the state entails that it is conscious, and is not essentially conscious just in case this is not so.

Sensations are good candidates for states that are essentially conscious. If an agent is in pain, this state would seem to be conscious. (This is not to deny that the agent might fail to attend to it.) Beliefs, knowledge, and other cognitive states are good candidates for states that might be conscious but not essentially so. One may truly say that an agent knows the rules of his language even though this knowledge is unconscious. Perception presents a hard case, as is demonstrated by the phenomenon of blindsight, in which subjects report that they do not see anything in portions of the visual field and yet their performance on forced-choice tasks suggests otherwise (Weiskrantz 1986). Clearly *some* information processing is going on in such cases, but it is not obvious that what is going on is properly described as perception, or at least as perceptual experience. It is plausible to suppose that indecision about how to describe matters here derives in part from indecision about whether perceptual states or experiences are essentially conscious.

Transitive vs. Intransitive Consciousness

In the case of creature consciousness, one may speak of someone's being conscious *simpliciter* and of someone's being conscious *of* something or other. Malcolm calls the first "intransitive" and the second "transitive" consciousness (e.g., Armstrong and Malcolm 1984). To say that a person is conscious *simpliciter* is a way of saying that the person is awake or alert. So the study of creature intransitive consciousness may be assimilated to the study of what it is to be alert. The denial of consciousness *simpliciter* does not entail a denial of psychological states altogether. If an agent is fast asleep on a couch, one may truly say both that the agent is unconscious *and* that the agent believes that snow is white. Humphrey (1992) speculates that the *notion* of intransitive consciousness is a recent one, perhaps about two hundred years old, but presumably people *were* on occasion intransitively conscious (i.e., alert or awake) prior to that date.

To say that a person is conscious of something seems to be a way of saying that the person knows or has beliefs about that thing. To say that one is conscious of a noise overhead is to say that one knows there is a noise overhead, though perhaps with the accompanying implication that one knows this only vaguely. So the study of creature transitive consciousness may be assimilated to the study of knowledge or beliefs. It is sometimes suggested that "consciousness" and "awareness" are synonyms. This is true only on the assumption that what is intended is creature transitive consciousness, since awareness is always *by* someone *of* something.

Intentional vs. Nonintentional Consciousness

While the transitive/intransitive distinction has no obvious analogue in the case of state consciousness—a state is not *itself* awake or alert, nor is it aware of anything—a related distinction is that between intentional and nonintentional conscious states. An intentional conscious state is *of* something in the sense that it represents the world as being a certain way—such states exhibit "intentionality," to adopt the traditional word. A nonintentional conscious state is a state that does not represent the world as being in some way. It is sometimes suggested that bodily sensations (itches, pains) are states of this second kind, while perceptual experiences (seeing a blue square on a red background) are cases of the first. But the matter is controversial given that to have a pain in one's foot seems to involve among other things representing one's foot as being in some condition or other, a fact that suggests that even here there is an intentional element in consciousness (see Intentionality).

Phenomenal vs. Access Consciousness

Block (1995) distinguishes two kinds of state consciousness: **phenomenal consciousness** and access consciousness. The notion of a phenomenally conscious state is usually phrased in terms of "What is it like ... ?" (e.g., Nagel 1974, "What is it like to be a bat?"). For a state to be phenomenally conscious is for there to be something it is akin to being, in that state. In the philosophical literature, the terms "qualia," "phenomenal character," and "experience" are all used as rough synonyms for phenomenal consciousness in this sense, though unfortunately there is no terminological consensus here.

For a state to be **access conscious** is, roughly, for the state to control rationally, or be poised to control rationally, thought and behavior. For creatures who have language, access consciousness closely correlates with reportability, the ability to express the state at will in language. Block suggests, among other things, that a state can be phenomenally conscious without its being access conscious.

For example, suppose one is engaged in intense conversation and becomes aware only at noon that there is a jackhammer operating outside—at five to twelve, one's hearing the noise is phenomenally conscious, but it is not access conscious. Block argues that many discussions both in the sciences and in philosophy putatively about phenomenal consciousness are in fact about access consciousness. Block's distinction is related to one made by Armstrong (1980) between experience and consciousness. For Armstrong, consciousness is attentional: A state is conscious just in case one attends to it. However, since one can have an experience without attending to it, it is possible to divorce experience and consciousness in this sense.

Within the general concept of access consciousness, a number of different strands may be distinguished. For example, an (epistemologically) normative interpretation of the notion needs to be separated from a nonnormative one. In the former case, the mere fact that one is in an access-conscious state puts one in a position to *know* or *justifiably* believe that one is; in the latter, being in an access-conscious state prompts one to think or believe that one is—here there is no issue of epistemic appraisal (see Epistemology). Further, an actualist interpretation of the notion needs to be separated from a counterfactual one. In the former case, what is at issue is whether one *does* know or think that one is in the state; in the latter, what is at issue is whether one *would*, provided other cognitive conditions were met. Distinguishing these various notions leads naturally into other issues. For example, consider the claim that it is essentially true of all psychological states that if one is in them, one would know that one is, provided one reflects and has the relevant concepts. That is one way of spelling out the Cartesian idea (recently defended by Searle [1992]) that the mind is "transparent" to itself.

There are hints both in Block (1995) and in related discussion (e.g., Davies and Humphreys 1993) that the phenomenal/access distinction is in (perhaps rough) alignment with both the intentional/nonintentional and the essential/nonessential distinction. The general idea is that phenomenally conscious states are *both* essentially so *and* nonintentional, while access-conscious states are neither. In view of the different interpretations of access consciousness, however, it is not clear that this is so. Psychological states might well be both essentially access conscious and phenomenally conscious. And, as indicated earlier, perhaps all conscious states exhibit intentionality in some form or other.

Self-consciousness

Turning back from state consciousness to creature consciousness, a notion of importance here is self-consciousness, i.e., one's being conscious of oneself as an agent or self or (in some cases) a person. If to speak of a creature's being 'conscious' of something is to speak about knowledge or beliefs, to attribute self-consciousness is to attribute to a creature knowledge or beliefs that the creature is a self or an agent. This would presumably require significant psychological complexity and perhaps cultural specificity. Proposals like those of Jaynes (1976) and Dennett (1992)—that consciousness is a phenomenon that emerges only in various societies—are best interpreted as concerning self-consciousness in this sense, which becomes more natural the more one complicates the underlying notion of self or agent. (Parallel remarks apply to any notion of group consciousness, assuming such a notion could be made clear.)

Problems of Consciousness

If these are the varieties of consciousness, it is easy enough to say in general terms what the problems of consciousness are, i.e., to explain or understand consciousness in all its varieties. But demands for explanation mean different things to different people, so the matter requires further examination.

To start with, one might approach the issue from an unabashedly scientific point of view. Consciousness is a variegated phenomenon that is a pervasive feature of the mental lives of humans and other creatures. It is desirable to have an explanation of this phenomenon, just as it is desirable to have an explanation of the formation of the moon, or the origin of HIV/AIDS. Questions that might be raised in this connection, and indeed have been raised, concern the relation of consciousness to neural structures (e.g., Crick and Koch 1998), the evolution of consciousness (e.g., Humphrey 1992), the relation of consciousness to other psychological capacities (e.g., McDermott 2001), the relation of consciousness to the physical and social environment of conscious organisms (e.g., Barlow 1987), and relations of unity and difference among conscious states themselves (e.g., Bayne and Chalmers 2003).

The attitude implicit in this approach—that consciousness might be studied like other empirical phenomena—is attractive, but there are at least four facts that need to be confronted before it can be completely adopted:

1. No framework of ideas has as yet been worked out within which the study of consciousness can proceed. Of course, this does not exclude the possibility that such a framework might be developed in the future, but it does make the specific proposals (such as that of Crick and Koch 1998) difficult to evaluate.
2. In the past, one of the main research tasks in psychology and related fields was to study psychological processes that were not conscious. This approach yielded a number of fruitful lines of inquiry—e.g., Chomsky's (1966) idea that linguistic knowledge is to be explained by the fact that people have unconscious knowledge of the rules of their language—but will presumably have to be abandoned when it comes to consciousness itself.
3. As Block (1995) notes, the standard concept of consciousness seems to combine a number of different concepts, which in turn raises the threat that consciousness in one sense will be confused with consciousness in another.
4. The issue of consciousness is often thought to raise questions of a philosophical nature, and this prompts further questions about whether a purely scientific approach is appropriate.

What are the philosophical aspects of the issue of consciousness? In the history of the subject, the issue of consciousness is usually discussed in the context of another, the traditional mind–body problem. This problem assumes a distinction between two views of human beings: the materialist or physicalist view, according to which human beings are completely physical objects; and the dualist view, according to which human beings are a complex of both irreducibly physical and irreducibly mental features. Consciousness, then, emerges as a central test case that decides the issue. The reason is that there are a number of thought experiments that apparently make it plausible to suppose that consciousness is distinct from anything physical. An example is the **inverted spectrum hypothesis**, in which it is imagined that two people might be identical except for the fact that the sensation provoked in one when looking at blood is precisely the sensation provoked in the other when looking at grass (Shoemaker 1981). If this hypothesis represents a genuine possibility, it is a very short step to the falsity of physicalism, which, setting aside some complications, entails that if any two people are identical physically, they are identical psychologically. On the other hand, if the inverted spectrum is possible, then two people identical physically may yet differ in respect of certain aspects of their conscious experience, and so differ psychologically. In short, physicalists are required to argue that the inverted spectrum hypothesis does not represent a genuine possibility. And this places the issue of consciousness at the heart of the mind–body problem.

In contemporary philosophy of mind, the traditional mind–body problem has been severely criticized. First, most contemporary philosophers do not regard the falsity of physicalism as a live option (Chalmers [1996] is an exception), so it seems absurd to debate something one already assumes to be true. Second, some writers argue that the very notions within which the traditional mind–body problem is formed are misguided (Chomsky 2000). Third, there are serious questions about the legitimacy of supposing that reflection of possible cases such as the inverted spectrum could even in principle decide the question of dualism or materialism, which are apparently empirical, contingent claims about the nature of the world (Jackson 1998).

As a result of this critique, many philosophers reject the mind–body problem in its traditional guise. However, it is mistaken to infer from this that concern with the inverted spectrum and related ideas has likewise been rejected. Instead, the theoretical setting of these arguments has changed. For example, in contemporary philosophy, the inverted spectrum often plays a role not so much in the question of whether physicalism is true, but rather in questions about whether phenomenal consciousness is in principle irreducible or else lies beyond the limits of rational inquiry. The impact of philosophical issues of this kind on a possible science of consciousness is therefore straightforward.

In the philosophical debates just alluded to, the notion of consciousness at issue is phenomenal consciousness. In other areas of philosophy, other notions are more prominent. In epistemology, for example, an important question concerns the intuitive difference between knowledge of one's own mental states—which seems in a certain sense privileged or direct—and knowledge of the external world, including the minds of others. This question has been made more acute by the impact of externalism, the thesis that one's psychological states depend constitutively on matters external to the subject, factors for which direct knowledge is not plausible (Davies 1997). Presumably these issues will be informed by the study of access consciousness. Similarly, in discussions of the notion of personal identity and related questions about how and why persons are objects of special moral concern, the notion of self-consciousness plays an important

role. One might also regard both access consciousness and self-consciousness as topics for straightforward scientific study.

Prospects for Progress

Due to the influence of positivist and postpositivist philosophy of science in the twentieth century, it was at one time common to assume that some or all questions of consciousness were pseudo-questions. Recently it has been more common to concede that the questions are real enough. But what are the chances of progress here?

In light of the multifariousness of the issues, a formulaic answer to this question would be inappropriate. Access consciousness seems to be a matter of information processing, and there is reason to suppose that such questions might be addressed using contemporary techniques. Hence, many writers (e.g., Block 1995) find grounds for cautious optimism here, though this might be tempered depending on whether access consciousness is construed as involving a normative element. In the case of self-consciousness, the issue of normativity is also present, and there is the added complication that self-consciousness is partly responsive to questions of social arrangements and their impact on individual subjects.

But it is widely acknowledged that the hardest part of the issue is phenomenal consciousness. Here the dominant strategy has been an indirect one of attempting to reduce the overall number of problems. One way to implement this strategy is to attempt to explain the notion of phenomenal consciousness in terms of another notion, say, access consciousness or something like it. Some philosophers suggest that puzzlement about phenomenal consciousness is a cognitive illusion, generated by a failure to understand the special nature of concepts of phenomenal consciousness, and that once this puzzlement is dispelled, the way will be clear for a straightforward identification of phenomenal and access consciousness (e.g., Tye 1999). Others argue that discussions of phenomenal consciousness neglect the extent to which conscious states involve intentionality, and that once this is fully appreciated there is no bar to adopting the view that phenomenal consciousness is just access consciousness (e.g., Carruthers 2000).

The attractive feature of these ideas is that if successful, they represent both philosophical and scientific progress. But the persistent difficulty is that the proposed explanations are unpersuasive. It is difficult to rid oneself of the feeling that what is special about concepts of phenomenal consciousness derives from only what it is that they are concepts of, and this makes it unlikely that the puzzles of phenomenal consciousness are an illusion. And, while it is plausible that phenomenally conscious states are intentional, emphasizing this fact will not necessarily shed light on the issue, for the intentionality of phenomenal consciousness might be just as puzzling as phenomenal consciousness itself.

However, even if one agrees that phenomenal consciousness represents a phenomenon distinct from these other notions, and therefore requires a separate approach, there is still a way in which one might seek to implement the strategy of reducing the number of problems, for, as noted earlier, phenomenal consciousness is thought to present both a philosophical and a scientific challenge. But what is the relation between these two issues? It is common to assume that the philosophical problem needs to be removed before one can make progress on the science. But perhaps the reverse is true. If the philosophical problems can be seen to be a reflection partly of ignorance in the scientific domain, there is no reason to regard them as a further impediment to scientific study. This might not seem like much of an advance. But the study of consciousness has been hampered by the feeling that it presents a problem of a different order from more straightforward empirical problems. In this context, to combat that assumption is to move forward, though slowly.

DANIEL STOLJAR

References

Armstrong, D. (1980), "What Is Consciousness?" in *The Nature of Mind and Other Essays*. Ithaca, NY: Cornell University Press, 55–67.

Armstrong, D., and N. Malcolm (1984), *Consciousness and Causality*. Oxford: Blackwell.

Barlow, H. (1987), "The Biological Role of Consciousness," in C. Blakemore and S. Greenfield (eds.), *Mindwaves*. Oxford: Basil Blackwell.

Bayne, T., and D. Chalmers (2003), "What Is the Unity of Consciousness?" in A. Cleeremans (ed.), *The Unity of Consciousness: Binding, Integration and Dissociation*. Oxford: Oxford University Press.

Block, N. (1995), "On a Confusion About a Function of Consciousness," *Behavioral and Brain Sciences* 18: 227–247.

Chalmers, D. (1996), *The Conscious Mind*. New York: Oxford University Press.

Chomsky, N. (1966), *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

——— (2000), *New Horizons in the Study of Mind and Language*. Cambridge: Cambridge University Press.

Crick, F., and C. Koch (1990), "Towards a Neurobiological Theory of Consciousness," *Seminars in the Neurosciences* 2: 263–275.

Carruthers, P. (2000), *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge: Cambridge University Press.
Davies, M. (1997), "Externalism and Experience," in N. Block, O. Flanagan, and G. Guzeldere (eds.), *The Nature of Consciousness: Philosophical Debates*. Cambridge, MA: MIT Press, 309–328.
Davies, M., and G. Humphreys (1993), "Introduction," in M. Davies and G. Humphreys (eds.), *Consciousness: Psychological and Philosophical Essays*. Oxford: Blackwell.
Dennett , D. (1992), *Consciousness Explained*. Boston: Little, Brown and Co.
Dretske, D. (1993), "Conscious Experience," *Mind* 102: 263–283.
Humphrey, N. (1992), *A History of the Mind*. New York: Simon and Schuster.
Jackson, J. (1998), *From Metaphysics to Ethics*. Oxford: Clarendon.
Jaynes, J. (1976), *The Origin of Consciousness in the Breakdown of the Bicameral Mind*. Boston: Houghton Mifflin Co.
McDermott, D. (2001), *Mind and Mechanism*. Cambridge, MA: MIT Press.
Nagel, T. (1974), "What Is It Like to Be a Bat?" *Philosophical Review* 83: 7–22.
Rosenthal, D. (1992), "A Theory of Consciousness," in N. Block, O. Flanagan, and G. Guzeldere (eds.), *The Nature of Consciousness: Philosophical Debates*. Cambridge, MA: MIT Press, 729–754.
Searle, R. (1992), *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.
Shoemaker, S. (1981), "The Inverted Spectrum," *Journal of Philosophy* 74: 357–381.
Tye, M. (1999), "Phenomenal Consciousness: The Explanatory Gap as a Cognitive Illusion," *Mind* 108: 432.
Weiskrantz, L. (1986), *Blindsight*. Oxford: Oxford University Press.

See also **Cognitive Science; Connectionism; Intentionality; Psychology, Philosophy of**

CONSERVATION BIOLOGY

Conservation biology emerged in the mid-1980s as a science devoted to the conservation of biological diversity, or *biodiversity*. Its emergence was precipitated by a widespread concern that anthropogenic development, especially deforestation in the tropics (Gómez-Pompa, Vázquez-Yanes, and Guevera 1972), had created an extinction crisis: a significant increase in the rate of species extinction (Soulé 1985). From its beginning, the primary objective of conservation biology was the design of **conservation area networks** (CANs), such as national parks, nature reserves, and managed-use zones that protect areas from anthropogenic transformation.

Conservation biology, then, is a normative discipline, in that it is defined in terms of a practical goal in addition to the accumulation of knowledge about a domain of nature. In this respect, it is analogous to medicine (Soulé 1985). Like medicine, conservation biology performs its remedial function in two ways: through intervention (for example, when conservation plans must be designed for species at risk of imminent extinction) and through prevention (for example, when plans are designed to prevent decline in species numbers long before extinction is imminent).

The normative status of conservation biology distinguishes it from ecology, which is not defined in terms of a practical goal (see Ecology). Moreover, besides using the models and empirical results of ecology, conservation biology also draws upon such disparate disciplines as genetics, computer science, operations research, and economics in designing and implementing CANs. Each of these fields contributes to a comprehensive framework that has recently emerged about the structure of conservation biology (see "The Consensus Framework" below).

Different views about the appropriate target of conservation have generated distinct methodologies within conservation biology. How 'biodiversity' is defined and, correspondingly, what conservation plans are designed will partly reflect ethical views about what features of the natural world are valuable (Norton 1994; Takacs 1996). There exists, therefore, a close connection between conservation biology and environmental ethics.

The Concept of Biodiversity

'Biodiversity' is typically taken to refer to diversity at all levels of biological organization: molecules,

cells, organisms, species, and communities (e.g., Meffe and Carroll 1994). This definition does not, however, provide insight into the fundamental goal of conservation biology, since it refers to all biological entities (Sarkar and Margules 2002). Even worse, this definition does not exhaust all items of biological interest that are worth preserving: Endangered biological phenomena, such as the migration of the monarch butterfly, are not included within this definition (Brower and Malcolm 1991). Finally, since even a liberally construed notion of biodiversity does not capture the ecosystem processes that sustain biological diversity, some have argued that a more general concept of biological *integrity*, incorporating both diversity of entities and the ecological processes that sustain them, should be recognized as the proper focus of conservation biology (Angermeier and Karr 1994).

In response to these problems, many conservation biologists have adopted a pluralistic approach to biodiversity concepts (Norton 1994; Sarkar and Margules 2002; see, however, Faith [2003], who argues that this ready acceptance of pluralism confuses the [unified] concept of biodiversity with the plurality of different conservation strategies). Norton (1994), for example, points out that any measure of biodiversity presupposes the validity of a specific model of the natural world. The existence of several equally accurate models ensures the absence of any uniquely correct measure. Thus, he argues, the selection of a biodiversity measure should be thought of as a normative political decision that reflects specific conservation values and goals. Sarkar and Margules (2002) argue that the concept of biodiversity is implicitly defined by the specific procedure employed to prioritize places for conservation action (see ''Place Prioritization'' below). Since different contexts warrant different procedures, biodiversity should similarly be understood pluralistically.

Two Perspectives

Throughout the 1980s and early 90s, the discipline was loosely characterized by two general approaches, which Caughley (1994) described as the ''small-population'' and ''declining-population'' paradigms of conservation. Motivated significantly by the legal framework of the Endangered Species Act of 1973 and the National Forest Management Act of 1976, the **small-population paradigm** originated and was widely adopted in the United States (Sarkar 2005). It focused primarily on individual species threatened by extinction. By analyzing their distributions and habitat requirements, conservation biologists thought that ''minimum viable populations'' could be demonstrated, that is, population sizes below which purely stochastic processes would significantly increase the probability of extinction (see ''Viability Analysis'' below). It quickly became clear, however, that this methodology was inadequate. It required much more data than could be feasibly collected for most species (Caughley 1994) and, more importantly, failed to consider the interspecies dynamics essential to most species' survival (Boyce 1992).

Widely followed in Australia, the **declining-population paradigm** focused on deterministic, rather than stochastic, causes of population decline (Sarkar 2005). Unlike the small-population paradigm, its objective was to identify and eradicate these causes before stochastic effects became significant. Since the primary cause of population decline was, and continues to be, habitat loss, conservation biologists, especially in Australia, became principally concerned with protecting the full complement of regional species diversity and required habitat within CANs. With meager monetary resources for protection, conservation biologists concentrated on developing methods that identified representative CANs in minimal areas (see ''Place Prioritization'' below).

The Consensus Framework

Recently, a growing consensus about the structure of conservation biology has emerged that combines aspects of the small-population and declining-population paradigms into a framework that emphasizes the crucial role computer-based place prioritization algorithms play in conservation planning (Margules and Pressey 2000; Sarkar 2005). The framework's purpose is to make conservation planning more systematic, thereby replacing the ad hoc reserve design strategies often employed in real-world planning in the past. It focuses on ensuring the adequate representation and persistence of regional biodiversity within the socioeconomic and political constraints inherent in such planning.

Place Prioritization

The first CAN design methods, especially within the United States, relied almost exclusively on island biogeography theory (MacArthur and Wilson 1967) (see Ecology). It was cited as the basis for geometric design principles intended to minimize

extinction rates in CANs as their ambient regions were anthropogenically transformed (Diamond 1975). Island biogeography theory entails that the particular species composition of an area is constantly changing while, at an equilibrium between extinction and immigration, its species richness remains constant. The intention behind design principles inspired by the theory, therefore, was to ensure persistence of the maximum number of species, not the specific species the areas currently contained. However, incisive criticism of the theory (Gilbert 1980; Margules, Higgs, and Rafe 1982) convinced many conservation biologists, especially in Australia, that *representation* of the *specific species* that areas now contain, rather than the *persistence* of the *greatest number of species* at some future time, should be the first goal of CAN design. Computer-based place prioritization algorithms supplied a defensible methodology for achieving the first goal and an alternative to the problematic reliance on island biogeography theory in CAN design.

Place prioritization involves solving a resource allocation problem. Conservation funds are usually significantly limited and priority must be given to protecting some areas over others. The Expected Surrogate Set Covering Problem (ESSCP) and the Maximal Expected Surrogate Covering Problem (MESCP) are two prioritization problems typically encountered in biodiversity conservation planning (Sarkar 2004). Formally, consider a set of individual places called cells $\{c_j : j = 1, \ldots, n\}$; cell areas $\{a_j : j = 1, \ldots, n\}$; biodiversity surrogates $\Lambda = \{s_i : i = 1, \ldots, m\}$; representation targets, one for each surrogate $\{t_i : i = 1, \ldots, m\}$; probabilities of finding s_i at c_j $\{p_{ij} : i = 1, \ldots, m;\ j = 1, \ldots, n\}$; and two indicator variables $X_j (j = 1, \ldots, n)$ and $Y_i (i = 1, \ldots, m)$ defined as follows:

$$X_j = \begin{cases} 1, & \text{if } c_j \in \Gamma; \\ 0, & \text{otherwise;} \end{cases}$$

$$Y_i = \begin{cases} 1, & \text{if } \sum_{c_j \in \Gamma} p_{ij} > t_i; \\ 0, & \text{otherwise.} \end{cases}$$

ESSCP is the problem:

$$\text{Minimize } \sum_{j=1}^{n} a_j X_j \text{ such that } \sum_{j=1}^{n} X_j p_{ij} \geq t_i \text{ f}$$

$$\text{or } \forall s_i \in \Lambda.$$

Informally, find the set of cells Γ with the smallest area such that every representation target is satisfied. MESCP is the problem:

$$\text{Maximize } \sum_{i=1}^{m} Y_i \text{ such that } \sum_{j=1}^{n} X_j = \mathbf{M},$$

where $\mathbf{M}$ is the number of protectable cells. Informally, given the opportunity to protect $\mathbf{M}$ cells, find those cells that maximize the number of representation targets satisfied.

The formal precision of these problems allows them to be solved computationally with heuristic or exact algorithms (Margules and Pressey 2000; Sarkar 2005). Exact algorithms, using mixed integer linear programming, guarantee optimal solutions: the smallest number of areas satisfying the problem conditions. Since ESSCP and MESCP are NP hard, however, exact algorithms are computationally intractable for practical problems with large datasets. Consequently, most research focuses on heuristic algorithms that are computationally tractable for large datasets but do not guarantee minimal solutions.

The most commonly used heuristic algorithms are "transparent," so called because the exact criterion by which each solution cell is selected is known. These algorithms select areas for incorporation into CANs by iteratively applying a hierarchical set of conservation criteria, such as rarity, richness, and complementarity. Rarity, for example, requires selecting the cell containing the region's most infrequently present surrogates, which ensures that endemic taxa are represented. The criterion most responsible for the efficiency of transparent heuristic algorithms is complementarity: Select subsequent cells that complement those already selected by adding the most surrogates not yet represented (Justus and Sarkar 2002). In policymaking contexts, transparency facilitates more perspicuous negotiations about competing land uses, which constitutes an advantage over nontransparent heuristic prioritization procedures such as those based on simulated annealing.

Methodologically, the problem of place prioritization refocused theoretical research in conservation biology from general theories to algorithmic procedures that require geographically explicit data (Sarkar 2004). In contrast to general theories, which abstract from the particularities of individual areas, place prioritization algorithms demonstrated that adequate CAN design critically depends upon these particularities.

Surrogacy

Since place prioritization requires detailed information about the precise distribution of a region's biota, devising conservation plans for specific areas would ideally begin with an exhaustive series of field surveys. However, owing to limitations of

time, money, and expertise, as well as geographical and sociopolitical boundaries to fieldwork, such surveys are usually not feasible in practice. These limitations give rise to the problem of discovering a (preferably small) set of biotic or abiotic land attributes (such as subsets of species, soil types, vegetation types, etc.) the precise distribution of which is realistically obtainable and that adequately represents biodiversity as such. This problem is referred to as that of finding **surrogates** or ''indicators'' for biodiversity (Sarkar and Margules 2002).

This challenge immediately gives rise to two important conceptual problems. Given the generality of the concept of biodiversity as such, the first problem concerns the selection of those entities that should be taken to represent concretely biodiversity in a particular planning context. These entities will be referred to as the *true surrogate*, or objective parameter, as it is the parameter that one is attempting to estimate in the field. Clearly, selection of the true surrogate is partly conventional and will depend largely on pragmatic and ethical concerns (Sarkar and Margules 2002); in most conservation contexts the true surrogate will typically be species diversity, or a species at risk. Once a set of true surrogates is chosen, the set of biotic and/or abiotic land attributes that will be tested for their capacity to represent the true surrogates adequately must be selected; these are *estimator surrogates*, or indicator parameters.

The second problem concerns the nature of the relation of representation that should obtain between the estimator and the true surrogate: What does 'representation' mean in this context and under what conditions can an estimator surrogate be said to represent adequately a true surrogate? Once the true and estimator surrogates are selected and a precise (operational) interpretation of representation is determined, the adequacy with which the estimator surrogate represents the true surrogate becomes an empirical question. (Landres, Verner, and Thomas [1988] discuss the import of subjecting one's choice of estimator surrogate to stringent empirical testing.)

Very generally, two different interpretations of representation have been proposed in the conservation literature. The more stringent interpretation is that the distribution of estimator surrogates should allow one to *predict* the distribution of true surrogates (Ferrier and Watson 1997). The satisfaction of this condition would consist in the construction of a well-confirmed model from which correlations between a given set of estimator surrogates and a given set of true surrogates can be derived. Currently such models have met with only limited predictive success; moreover, since such models can typically be used to predict the distribution of only one surrogate at a time, they are not computationally feasible for practical conservation planning, which must devise CANs to sample a wide range of regional biodiversity. Fortunately, the solution to the surrogacy problem does not in practice require predictions of true-surrogate distributions.

A less stringent interpretation of representation assumes only that the set of places prioritized on the basis of the estimator surrogates adequately captures true-surrogate diversity up to some specified target (Sarkar and Margules 2002). The question of the *adequacy* of a given estimator surrogate then becomes the following: If one were to construct a CAN that samples the full complement of estimator-surrogate diversity, to what extent does this CAN also sample the full complement of true-surrogate diversity? Several different quantitative measurements can be carried out to evaluate this question (Sarkar 2004). At present, whether adequate surrogate sets exist for conservation planning remains an open empirical question.

Viability Analysis

Place prioritization and surrogacy analysis help identify areas that *currently* represent biodiversity. This is usually not, however, sufficient for successful biodiversity conservation. The problem is that the biodiversity these areas contain may be in irreversible decline and unlikely to persist. Since principles of CAN design inspired by island biogeography theory were, by the late 1980s, no longer believed to ensure persistence adequately, attention subsequently turned, especially in the United States, to modeling the probability of population extinction. These principles became known as **population viability analysis** (PVA).

PVA models focus primarily on factors affecting small populations, such as inbreeding depression, environmental and demographic stochasticity, genetic drift, and spatial structure. Drift and inbreeding depression, for example, may reduce the genetic variation required for substantial evolvability, without which populations may be more susceptible to disease, predation, or future environmental change. PVA modeling has not, however, provided a clear understanding of the general import of drift and inbreeding depression to population persistence in nature (Boyce 1992). Unfortunately, this kind of problem pervades PVA. In a trenchant review, for instance, Caughley (1994) concluded that the predominantly theoretical work done thus far in PVA had not been adequately tested with field data and that many of the tests that had been done were seriously flawed.

Models used for PVA have a variety of different structures and assumptions (see Beisinger and McCullough 2002). As a biodiversity conservation methodology, however, PVA faces at least three general difficulties:

1. Precise estimation of parameters common to PVA models requires enormous amounts of quality field data, which are usually unavailable (Fieberg and Ellner 2000) and cannot be collected, given limited monetary resources and the imperative to take conservation action quickly. Thus, with prior knowledge being uncommon concerning what mechanisms are primarily responsible for a population's dynamics, PVA provides little guidance about how to minimize extinction probability.
2. In general, the results of PVA modeling are extremely sensitive to model structure and parameter values. Models with seemingly slightly different structure may make radically different predictions (Sarkar 2004), and different parameter values for one model may produce markedly different predictions, a problem exacerbated by the difficulties discussed under (1).
3. Currently, PVA models have been developed almost exclusively for single species (occasionally two) and rarely consider more than a few factors affecting population decline. Therefore, only in narrow conservation contexts focused on individual species would they potentially play an important role. Successful models that consider the numerous factors affecting the viability of *multiple*-species assemblages, which are and should be the primary target of actual conservation planning, are unlikely to be developed in the near future (Fieberg and Ellner 2000).

For these reasons and others, PVA has not thus far uncovered nontrivial generalities relevant to CAN design. Consequently, attention has turned to more pragmatic principles, such as designing CANs to minimize distance to anthropogenically transformed areas. This is not to abandon the important goals of PVA or the need for sound theory about biodiversity persistence, but to recognize the weaknesses of existing PVA models and the imperative to act now given significant threats to biodiversity.

Multiple-Criterion Synchronization

The implementation and maintenance of CANs inevitably take place within a context of competing demands upon the allocation and use of land. Consequently, successful implementation strategies should ideally be built upon a wide consensus among agents with different priorities with respect to that usage. Thus, practical CAN implementation involves the attempt to optimize the value of a CAN amongst several different criteria simultaneously. Because different criteria typically conflict, however, the term "synchronization" rather than "optimization" is more accurate. The multiple-constraint synchronization problem involves developing and evaluating procedures designed to support such decision-making processes.

One approach to this task is to "reduce" these various criteria to a single scale, such as monetary cost. For example, cost-benefit analysis has attempted to do this by assessing the amount an agent is willing to pay to improve the conservation value of an area. In practice, however, such estimates are difficult to carry out and are rarely attempted (Norton 1987). Another method, based on multiple-objective decision-making models, does not attempt to reduce the plurality of criteria to a single scale; rather it seeks merely to eliminate those feasible CANs that are suboptimal when evaluated according to all relevant criteria. This method, of course, will typically not result in the determination of a uniquely best solution, but it may be able to reduce the number of potential CANs to one that is small enough so that decision-making bodies can bring other implicit criteria to bear on their ultimate decision (see Rothley 1999 and Sarkar 2004 for applications to conservation planning).

JUSTIN GARSON AND JAMES JUSTUS

References

Angermeier, P. L., and J. R. Karr (1994), "Biological Integrity versus Biological Diversity as Policy Directives," *BioScience* 44: 690–697.

Beisinger, S. R., and D. R. McCullough (eds.) (2002), *Population Viability Analysis*. Chicago: University of Chicago Press.

Boyce, M. (1992), "Population Viability Analysis," *Annual Review of Ecology and Systematics* 23: 481–506.

Brower, L. P., and S. B. Malcolm (1991), "Animal Migrations: Endangered Phenomena," *American Zoologist* 31: 265–276.

Caughley, G. (1994), "Directions in Conservation Biology," *Journal of Animal Ecology* 63: 215–244.

Diamond, J. (1975), "The Island Dilemma: Lessons of Modern Biogeographic Studies for the Design of Natural Reserves," *Biological Conservation* 7: 129–146.

Faith, D. P. (2003), "Biodiversity," in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/sum2003/entries/biodiversity

Ferrier, S., and G. Watson (1997), *An Evaluation of the Effectiveness of Environmental Surrogates and Modelling*

Techniques in Predicting the Distribution of Biological Diversity: Consultancy Report to the Biodiversity Convention and Strategy Section of the Biodiversity Group, Environment Australia. Arimidale, New South Wales: Environment Australia.

Fieberg, J., and S. P. Ellner (2000), "When Is It Meaningful to Estimate an Extinction Probability?" *Ecology* 8: 2040–2047.

Gilbert, F. S. (1980), "The Equilibrium Theory of Island Biogeography: Fact or Fiction?" *Journal of Biogeography* 7: 209–235.

Gómez-Pompa, A., C. Vázquez-Yanes, and S. Guevera (1972), "The Tropical Rain Forest: A Nonrenewable Resource," *Science* 177: 762–765.

Justus, J., and S. Sarkar (2002), "The Principle of Complementarity in the Design of Reserve Networks to Conserve Biodiversity: A Preliminary History," *Journal of Biosciences* 27: 421–435.

Landres, P. B., J. Verner, and J. W. Thomas (1988), "Ecological Uses of Vertebrate Indicator Species: A Critique," *Conservation Biology* 2: 316–328.

MacArthur, R., and E. O. Wilson (1967), *The Theory of Island Biogeography*. Princeton, NJ: Princeton University Press.

Margules, C. R., and R. L. Pressey (2000), "Systematic Conservation Planning," *Nature* 405: 242–253.

Margules, C., A. J. Higgs, and R. W. Rafe (1982), "Modern Biogeographic Theory: Are There Lessons for Nature Reserve Design?" *Biological Conservation* 24: 115–128.

Meffe, G. K., and C. R. Carroll (1994), *Principles of Conservation Biology*. Sunderland, MA: Sinauer Associates.

Norton, B. G. (1987), *Why Preserve Natural Variety?* Princeton, NJ: Princeton University Press.

Norton, B. G. (1994), "On What We Should Save: The Role of Cultures in Determining Conservation Targets," in P. L. Forey, C. J. Humphries, and R. I. Vane-Wright (eds.), *Systematics and Conservation Evaluation*. Oxford: Clarendon Press, 23–29.

Rothley, K. D. (1999), "Designing Bioreserve Networks to Satisfy Multiple, Conflicting Demands," *Ecological Applications* 9: 741–750.

Sarkar, S. (2004), "Conservation Biology," in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. http://stanford.edu/entries/conservation-biology

——— (2005), *Biodiversity and Environmental Philosophy*. New York: Cambridge University Press.

Sarkar, S., and C. R. Margules (2002), "Operationalizing Biodiversity for Conservation Planning," *Journal of Biosciences* 27: 299–308.

Soulé, M. E. (1985), "What Is Conservation Biology?" *BioScience* 35: 727–734.

Takacs, D. (1996), *The Idea of Biodiversity: Philosophies of Paradise*. Baltimore: Johns Hopkins University Press.

CONSTRUCTIVE EMPIRICISM

See **Instrumentalism; Realism**

CONVENTIONALISM

Conventionalism is a philosophical position according to which the truth of certain propositions, such as those of ethics, aesthetics, physics, mathematics, or logic, is in some sense best explained by appeal to intentional human actions, such as linguistic stipulations. In this article, the focus will be on conventionalism concerning physics, mathematics, and logic. Conventionalism concerning empirical science and mathematics appears to have emerged as a distinctive philosophical position only in the latter half of the nineteenth century.

The philosophical motivations for conventionalism are manifold. Early conventionalists such as Pierre Duhem and Henri Poincaré, and more recently Adolf Grünbaum, have seen conventionalism about physical or geometrical principles as

justified in part by what they regard as the underdetermination of a physical or geometrical theory by empirical observation (see Duhem Thesis; Poincaré, Henri; Theories, Underdetermination of). An (empirically) arbitrary choice of a system from among empirically equivalent theories seems required.

A second motivation, also suggested by Duhem and Poincaré, and explicitly developed later by Rudolf Carnap, stems from their view that any description of the empirical world presupposes a suitable descriptive apparatus, such as a geometry, a metric, or a mathematics. In some cases there appears to be a choice as to which descriptive apparatus to employ, and this choice involves an arbitrary convention (see Carnap, Rudolf).

A third motivation for conventionalism, emphasized by philosophers such as Moritz Schlick, Hans Hahn, and Alfred Ayer, is that appeal to conventions provides a straightforward explanation of *a priori* propositional knowledge, such as knowledge of mathematical truths (see Ayer, Alfred Jules; Hahn, Hans; Schlick, Moritz).Such truths, it was thought, are known *a priori* in virtue of the fact that they are stipulated rather than discovered.

It is important to note that conventionalists do not regard the selection of a set of conventions to be wholly arbitrary, with the exception of the radical French conventionalist Edouard LeRoy, who did (see Giedymin 1982, 118–28). Conventionalists acknowledge that pragmatic or instrumental considerations involving human capacities or purposes might be relevant to the adoption of a set of conventions (see Instrumentalism). However, they insist that the empirical evidence does not compel one choice rather than another.

This article focuses on several major figures and issues, but there are a number of other important philosophers with broadly conventionalist leanings, including Pierre Duhem, Kazimierz Ajdukiewicz, and Ludwig Wittgenstein, that for the sake of brevity will not here receive the attention they deserve.

Poincaré and Geometric Conventionalism

The development of non-Euclidean geometries and subsequent research into the foundations of geometry provided both the inspiration and the model for much of the conventionalism of the early twentieth century (see Space-Time). The central figure in early conventionalism was the French mathematician and philosopher Henri Poincaré (see Poincaré, Henri). Although a Kantian in his philosophy of mathematics, Poincaré shared with many late-nineteenth-century mathematicians and philosophers the growing conviction that the discovery of non-Euclidean geometries, such as the geometries of Lobatschevsky and Riemann, conjoined with other developments in the foundations of geometry, rendered untenable the Kantian treatment of geometrical axioms as a form of synthetic *a priori* intuition.

The perceived failings of Kant's analysis of geometry did not, however, lead Poincaré to an empiricist treatment of geometry. If geometry were an experimental science, he reasoned, it would be open to continual revision or falsified outright (the perfectly invariable solids of geometry are never empirically discovered, for instance) (Poincaré [1905] 1952, 49–50). Rather, geometrical axioms are conventions. As such, the axioms (postulates) of a systematic geometry constitute *implicit definitions* of such primitive terms of the system as "point" and "line" (50). This notion of implicit definition originated with J. D. Gergonne (1818), who saw an analogy between a set of sentences with n undefined terms and a set of equations with n unknowns. The two are analogous in that the roots that satisfy the equations are akin to interpretations of the undefined terms in the set of sentences under which the sentences are true. Of course, not every set of equations determines a set of values, and so too not every set of sentences (system of axioms) constitutes an implicit definition of its primitive terms. Poincaré accommodated this fact by recognizing two constraints on the admissibility of a set of axioms. First, the set must be consistent, and second, it must uniquely determine the objects defined (1952, 150–3).

Although the axioms were in his view conventional, Poincaré thought that experience nonetheless plays a role within geometry. The genesis of geometrical systems is closely tied to experience, in that the systems of geometry that are constructed are selected on the basis both of prior idealized empirical generalizations and of the simplicity and convenience that particular systems (such as Euclidean geometry) may afford their users ([1905] 1952, 70–1; 1952, 114–5). But these empirical considerations do not provide a test of empirically applied geometries. Once elevated to the status of a convention, a geometrical system is not an empirical theory, but is rather akin to a language used, in part, to frame subsequent empirical assertions. As a system of implicit definitions, it is senseless to speak of one geometry being "more true" than another ([1905] 1952, 50).

Poincaré therefore rejected the supposition that an empirical experiment could compel the selection of one geometry over another, provided that both

satisfied certain constraints. He was inspired by Sophus Lie's (1959) group-theoretic approach to geometrical transformations, according to which, given an n-dimensional space and the possible transformations of figures within it, only a finite group of geometries with differing coordinate systems are possible for that space. Among these geometries there is no principled justification for selecting one (such as Plücker line geometry) over another (such as sphere geometry) (Poincaré [1905] 1952, 46–7). Indeed, one geometry within the group may be transformed into another given a certain method of translation, which Lie derived from Gergonne's theory of reciprocity. Poincaré conjoined the intertransformability of certain geometrical systems with the recognition that alternative geometries yield different conventions governing the notions of "distance" or "congruence" to argue that any attempt to decide by experiment between alternative geometries would be futile, since interpretations of the data presupposed geometric conventions. In a much-discussed passage he wrote:

> If Lobatchevsky's geometry is true, the parallax of a very distant star will be finite. If Riemann's is true, it will be negative. These are the results which seem within the reach of experiment .... But what we call a straight line in astronomy is simply the path of a ray of light. If, therefore, we were to discover negative parallaxes, or to prove that all parallaxes are higher than a certain limit, we should have a choice between two conclusions: we could give up Euclidean geometry, or modify the laws of optics, and suppose that light is not rigorously propagated in a straight line. ([1905] 1952, 72–3)

On one reading, advanced by Grünbaum, this passage affirms that there is no fact of the matter as to which is the "correct" metric, and hence supports the *conventionality of spatial metrics*. This reading is further discussed below. Another reading of this passage sees in it an argument closely akin to Duhem's argument leading to the denial of the possibility of a "crucial experiment" that could force the acceptance or elimination of a geometrical theory (see Duhem [1906] 1954, 180–90). Interestingly, Poincaré uses the argument to support a distinction between "conventional" truths and empirical truths, whereas Quine later adduces similar Duhemian considerations on behalf of his claim that there is no such distinction to be drawn (see Duhem Thesis; Quine, Willard Van Orman).

However he may have intended his parallax example, Poincaré would probably have accepted both the conventionality of the spatial metric and the absence of any experimental test capable of deciding between two alternative geometries. As a further illustration of these issues, he proposed a well-known thought experiment ([1905] 1952, 65–7). Imagine a world consisting of a large sphere subject to the following temperature law: At the center of the sphere the temperature is greatest, and at the periphery it is absolute zero. The temperature decreases uniformly as one moves toward the circumference in proportion to the formula $R^2 - r^2$, where R is the radius of the sphere and r the distance from the center. Assume further that all bodies in the sphere are in perfect thermal equilibrium with their environment, that all bodies share a coefficient of dilation proportional to their temperature, and that light in this sphere is transmitted through a medium whose index of refraction is $1/(R^2 - r^2)$. Finally, suppose that there are inhabitants of this sphere and that they adopt a convention that allows them to measure lengths with rigid rods. If the inhabitants assume that these rods are invariant in length under transport, and if they further triangulate the positions in their world with light rays, they might well come to the conclusion that their world has a Lobatchevskian geometry. But they could also infer that their world is Euclidean, by postulating the universal physical forces just described. Again, no empirical experiment seems adequate to establish one geometry over the other, since both are compatible with all known possible observations given appropriate auxiliary hypotheses. Rather, a conventional choice of a geometry seems called for. The choice will be motivated by pragmatic factors, perhaps, but not "imposed" by the experimental facts (Poincaré [1905] 1952, 70–1).

Grünbaum's Conventionalism: The Metric and Simultaneity

Poincaré's parallax and sphere-world examples motivated his view that the choice of a spatial metric is conventional. Adolph Grünbaum has defended this conventionalist conclusion at length (Grünbaum 1968, 1973). Grünbaum claims that there is no unique metric "intrinsic" to a spatial manifold, since between any two points on a real number line, there is an uncountably infinite continuum of points. Hence, if one wishes to specify a metric for a manifold, one must employ "extrinsic" devices such as measuring rods, and must further stipulate the rods' behavior under transport.

Grünbaum defends what he takes to be Poincaré's metric conventionalism against a variety of objections. Perhaps the most serious objection is the claim that Poincaré's result illustrates only a *trivial* point about the conventionality of referring expressions, an objection first directed against Poincaré by Arthur Eddington and subsequently

developed by Hilary Putnam and Paul Feyerabend. Eddington objected that Poincaré's examples illustrated not that metrical *relations* (such as the relations of equality or of the congruence of two rods) were conventional, but rather—and only—that the meanings of *words* such as "equal" and "congruent" were conventional. This, Eddington objected, was a trivial point about *semantical* conventionality, and the fact that the selection of different metrics could yield different but apparently equipollent geometries illustrated only that "the meaning assigned to length and distance has to go along with the meaning assigned to space" (Eddington 1953, 9–10). Eddington suggested, and Feyerabend later developed, a simple illustration of this point within the theory of gases. Upon the discovery that Boyle's law:

$$pv = RT$$

holds only approximately of real gases, one could *either* revise Boyle's law in favor of van der Waal's law:

$$(p + a/v^2)(v - b) = RT$$

or one could preserve Boyle's law by redefining pressure as follows:

$$\text{Pressure} =_{\text{Def}} (p + a/v^2)(1 - b/v).$$

(Feyerabend, quoted in Grnbaum 1973, 34)

The possibility of such a redefinition, the objection proceeds, is physically and philosophically uninteresting. Provided that a similar move is available in Poincaré's own examples with expressions like "congruent," Poincaré's examples appear to illustrate only platitudes about semantic conventionality.

Grünbaum tries to show that Poincaré's conventionalism is not merely an example of what Grünbaum calls "trivial semantic conventionality" by showing that the conventionality of the metric is the result of a certain absence of "intrinsic" structure within the space-time manifold, which structure can (or must) then be imposed "extrinsically":

> And the metric amorphousness of these continua then serves to *explain* that even *after* the word "congruent" has been pre-empted semantically as a spatial or temporal *equality* predicate by the axioms of congruence, congruence remains ambiguous in the sense that these axioms still allow an infinitude of mutually exclusive congruence classes of intervals.
>
> (Grünbaum 1973, 27)

Grünbaum developed a related argument to the effect that the simultaneity relation in special relativity (SR) involves a conventional element. A sympathetic reconstruction of the argument will be attempted here, omitting a number of details in order to present the gist of the argument as briefly and intuitively as possible. Within Newtonian mechanics (NM), there is no finite limit to the speed at which causal signals can be sent. Further, one might take it to be a defining feature of causal relations that effects cannot precede their causes. This asymmetry allows for a distinction between events that are temporally before or after a given event on a world line. There is then only one simultaneity relation within the space-time of NM meeting the constraints mentioned. Consider events (or space-time points, if one prefers) a and b on two distinct parallel world lines. Event a is simultaneous with b just in case a is not causally connectible to any event on the future part of b's world line but is causally connectible to any event on the past of b's world line. Within SR, however, there is an upper limit to the speed of causal signals. Thus the constraint that effects cannot precede their causes allows any one of a continuum of points along a parallel world line to be a candidate for simultaneity with a given event on the world line of an inertial observer. Grünbaum calls such points "topologically simultaneous" with a point o on the observer's world line. In claiming that simultaneity is (to some extent) conventional within the SR picture, Grünbaum is in good company, as Einstein makes a claim to this effect in his original 1905 paper. (It is important to distinguish the issue of the *conventionality* of simultaneity within SR, which has been controversial, from the (observer or frame) relativity of simultaneity within SR, which is unchallenged; see Space-Time.)

Grünbaum claims that the existence of a continuum of possible "planes of simultaneity" through a given point is explained by, or reflects, the fact that the simultaneity relation is not "intrinsic" to the manifold of events within special relativity. One objection that a number of Grünbaum's critics such as Putnam (in Putnam 1975b) have had to his treatment of conventionalism (concerning both the metric and the simultaneity relation) is that the notion of an intrinsic feature, which plays a crucial role for Grünbaum, is never adequately clarified. David Malament (1977) provides a natural reading of 'intrinsic': An intrinsic property or relation of a system is "definable" in terms of a set of basic features. The intuitive idea is that if a relation is definable from relations that are uncontroversially intrinsic, then these ought to count as intrinsic as well. Malament goes on to show that if one takes certain fairly minimal relations to

be given as intrinsic, then a unique simultaneity relation, in fact the standard one, is definable from them in a fairly well defined sense of 'definable. '

Malament's result has been taken by many to have definitively settled the issue of whether simultaneity is conventional within SR (see e.g., Norton 1992). In this view, the standard simultaneity relation is nonconventional, since it is definable, or "logically constructible," from uncontroversially intrinsic features of Minkowski space-time, the space-time of SR. Furthermore, standard simultaneity is the only such nonconventional simultaneity relation.

However, there are dissenters, including Grünbaum. A number of authors have attacked one or another of Malament's premises, including for instance the claim that simultaneity must be a transitive relation, that is, the requirement that for arbitrary events x, y, and z, if x sim y and y sim z, then x sim z. Sarkar and Stachel have shown that even if it is allowed that simultaneity must be an equivalence relation, other simultaneity relations (the backward and the forward light cones of events on the given world line) are definable from relations that Malament takes as basic or intrinsic (see Sarkar and Stachel 1999). Peter Spirtes (1981) shows that Malament's result is highly sensitive to the choice of basic or intrinsic relations, which fact might be taken to undercut the significance of Malament's result as well.

A conventionalist might raise a different sort of objection to Malament's results, questioning their relevance to a reasonable, although perhaps not the only reasonable, construal of the question as to whether simultaneity is conventional within SR. Suppose for simplicity that any relation that one is willing to call a simultaneity relation is an equivalence relation, that causes must always precede their effects, and that there is an upper bound to the speed of causal influences within SR. One may now ask whether more than one relation can be defined (extrinsically or otherwise) on the manifold of events (or "space-time points"), meeting these criteria within SR. That is, one might ask whether, given any model M of T (roughly, Minkowski space-time, the "standard" space-time of SR), there is a unique expansion of the model to a model M' for T', where T' adds to T sentences containing a symbol 'sim', which sentences in effect require that 'sim' be an equivalence relation and that causes are not 'sim' with their effects. It is a straightforward matter to see, as Grünbaum shows (the details are omitted here), that there is no such unique expansion of M. The constraints mentioned leave room for infinitely many possible interpretations of 'sim'. In this sense, the simultaneity relation (interpretation of 'sim') might naturally be said to be conventional within SR. In contrast, given a model for the causal structure of NM, the meaning constraints on the concept of simultaneity that are assumed here yield a unique expansion (a unique extension for 'sim').

It should be noted that the conventionality of simultaneity within SR and its nonconventionality within NM in the sense just described reflect structural differences between the two theories (or their standard models). Thus this form of conventionalism appears to escape the charge that the only sense in which simultaneity is conventional is the trivial sense in which the word "simultaneous" can be used to denote different relations. As Grünbaum frequently emphasizes, the interesting cases of conventionality arise only when one begins with an already meaningful term or concept, whose meaning is constrained in some nontrivial way. It also allows one to interpret Einstein as making a substantive, yet fairly obvious point when he claims that the standard simultaneity relation within SR is a conventional, or stipulated, choice.

How does this version of conventionalism relate to Malament's arguments? It will be helpful to note a puzzle before proceeding. Wesley Salmon (1977) argues at length that the one-way speed of light is not empirically identifiable within SR. (The basic problem is that the one-way speed seems measurable only by using distant synchronized clocks, but the synchronization of distant clocks would appear to involve light signals and presuppositions about their one-way speeds.) This fact appears to yield as an immediate consequence the conventionality (in the sense of empirical admissibility described above) of simultaneity within SR, AU:42 given the various apparently admissible ways of synchronizing distant clocks within a reference frame. The puzzle is this: On one hand, the standard view has it that Malament's result effectively proves the falsehood of conventionalism (about simultaneity within SR); on the other hand, Salmon's arguments appear to entail the truth of conventionalism. Yet, his arguments for the empirical inaccessibility of the one-way speed of light seem sound. No one has convincingly shown what is wrong with them. All that the nonconventionalist seems able to provide is the indirect argument that Malament is right, so Salmon (and Grünbaum) must be wrong.

In the present analysis, this puzzle is dissolved by distinguishing two proposals, each of which may justifiably be called versions of conventionalism.

One version claims that a *relation* (thought of as a set or extension) is conventional within a model for a theory if it is not intrinsic within M, where 'intrinsic' might then be interpreted as definability (in the sense of logical constructibility) from uncontroversially intrinsic relations within M. The other version claims that the extension of a *concept* is conventional within a model (or theory) if the meaning of the concept does not entail a unique extension for the concept within the model (or within all models of the theory). Call these versions of conventionalism C_1 and C_2, respectively. Malament shows that simultaneity is not C_1 conventional within SR (leaving aside other possible objections to his result). Grünbaum and Salmon show that simultaneity is C_2 conventional within SR. Malament further considers relations definable from the causal connectibility relation (together with the world line of an inertial observer) within SR, and notes that only one is a nontrivial candidate simultaneity relation (relative to the corresponding inertial frame). He shows there is only one, and that therefore there is a unique intrinsic (and hence non–C_1 conventional) simultaneity relation within SR, the standard one. The C_2 conventionalist may respond that although Malament has pointed out an interesting feature of a particular candidate interpretation of the already meaningful term "simultaneous," he has not thereby ruled out all other candidate interpretations. One should not, the C_2 defender will argue, confuse the metalinguistic notion 'definable within M (or T)' with a meaning constraint on our concept of simultaneity, i.e., a constraint on potential interpretations of our already meaningful term "simultaneous." It would be absurd to claim that what was meant all along by 'simultaneous' required not only that any candidate be an equivalence relation and that it must "fit with" causal relations in that effects may not precede their causes, but also that for any two events, they are simultaneous if and only if they are in a relation that is definable from the causal connectibility relation.

A possible interpretation of the debate concerning the conventionality of simultaneity within SR, on the present construal, is that Grünbaum, perhaps in order to avoid the charge of simply rediscovering a form of trivial semantic conventionality, appealed to his notion of an intrinsic relation. Failure to express or denote an intrinsic feature was then said to characterize the interesting (nontrivial) cases of conventionality. This notion led to a number of difficulties for Grünbaum, culminating in Malament's apparent refutation of the conventionality of simultaneity within SR. However, it has been argued above that the appeal to intrinsicness turns out to be unnecessary in order to preserve a nontrivial form of conventionalism, C_2.

Grünbaum's arguments concerning the conventionality of the space-time metric for continuous manifolds involve more complex issues than those in the simultaneity case. But some key elements are common to both disputes. Grünbaum claims that continuous manifolds do not have intrinsic metrics, and critics complain that the notion of an intrinsic feature is unclear (see e.g., Stein 1977). Some of the claims that Grünbaum makes lend support to the idea that he is concerned with C_1 conventionalism. He claims, for example, that for a discrete manifold (such that between any two points there are only finitely many points), there is an intrinsic and hence nonconventional metric, where the "distance" between any two points is the number of points between them, plus one. This claim makes sense to the extent that one is concerned with C_1, that is, with the question of whether a relation is definable or logically constructible from a basic set. Grünbaum's arguments do not seem well suited for showing that such a metric would be C_2 nonconventional, since nothing about the meaning of 'distance' constrains the adoption of this as opposed to infinitely other metrics.

Another worry that might be raised concerns Grünbaum's insistence that distance functions defined in terms of physical bodies and their behaviors under transport are extrinsic to a manifold. If the manifold is a purely mathematical object, then such a view seems more plausible. But if one is concerned with a physical manifold of space-time points, it is less obvious that physical bodies should be treated as extrinsic. More importantly, even if one grants Grünbaum that physical bodies and their behaviors are extrinsic to the physical space-time manifold, one might want to grant a scientist the right to specify which features of the structures *that are being posited* are to count as basic or intrinsic. For example, Grünbaum proposes that Newton in effect made a conceptual error in claiming that whether the temporal interval between one pair of instants is the same as that between another pair of instants is a factual matter, determined by the structure of time itself. Since instants form a temporal continuum in Newton's picture, there is, according to Grünbaum, no intrinsic temporal metric, contrary to Newton's claims. However, one can imagine a Newtonian responding with bewilderment. Does Newton not get to say what relations are intrinsic to the structures that he is positing?

It is difficult to see how to make room for a notion of intrinsicness that can do all of the philosophical work that Grünbaum requires of it, as Stein (1977) and others argue at length. On the other hand, Grünbaum's conventionalism about simultaneity within SR remains defensible (in particular, it escapes the trivialization charge as well as Malament's attempted refutation) if his conclusions are interpreted in the sense of C_2 conventionalism described above.

Mathematical and Logical Conventionalism

Up to now this discussion has focused on conventionalist claims concerning principles of an empirical science, physics. Other propositions that have attracted conventionalist treatment are those from the nonempirical sciences of mathematics and logic. Conventionalism seems especially attractive here, particularly to empiricists and others who find the notion of special faculties of mathematical intuition dubious but nevertheless wish to treat known mathematical or logical propositions as nonempirical.

The axiomatic methods that had motivated Poincaré's geometric conventionalism discussed above received further support with the publication of David Hilbert's *Foundations of Geometry* ([1921] 1962). Hilbert disregarded the intuitive or ordinary meanings of such constituent terms of Euclidean geometry as "point," "line," and "plane," and instead proposed regarding the axioms as purely formal posits (see Hilbert, David). In other words, the axioms function as generalizations about whatever set of things happens to satisfy them. In addition to allowing Hilbert to demonstrate a number of significant results in pure geometry, this method of abstracting from particular applications had immediate philosophical consequences. For instance, in response to Gottlob Frege's objection that without fixing a reference for primitive expressions like "between," Hilbert's axiomatic geometry would be equivocal, Hilbert wrote:

> But surely it is self-evident that every theory is merely a framework or schema of concepts together with their necessary relations to one another, and that the basic elements can be construed as one pleases. If I think of my points as some system or other of things, e.g., the system of love, of law, or of chimney sweeps ... and then conceive of all my axioms as relations between these things, then my theorems, e.g., the Pythagorean one, will hold of these things as well.
>
> (Hilbert 1971, 13)

Moritz Schlick was quick to find in Hilbert's conclusions the possibility of generalizing conventionalism beyond geometry (see Schlick, Moritz).

If the reference of the primitive concepts of an axiomatic system could be left undetermined, as Hilbert had apparently demonstrated, then the axioms would be empty of empirical content, and so knowledge of them would appear to be *a priori*. Like Poincaré, Schlick rejected a Kantian explanation of how such knowledge is possible, and he saw in Hilbert's approach a demonstration of how an implicit definition of concepts could be obtained through axioms whose validity had been guaranteed (Schlick [1925] 1985, 33). In his *General Theory of Knowledge* Schlick distinguished explicitly between "ordinary concepts," which are defined by ostension, and implicit definitions. Of the latter, he wrote that

> [a] system of truths created with the aid of implicit definitions does not at any point rest on the ground of reality. On the contrary, it floats freely, so to speak, and like the solar system bears within itself the guarantee of its own stability. (37)

The guaranteed stability was to be provided by a consistency proof for a given system of axioms, which Schlick, like Poincaré and Hilbert before him, claimed was a necessary condition in a system of symbolism. Schlick followed Poincaré in treating the axioms as conventions, and hence as known *a priori* through stipulation (Schlick [1925] 1985, 71). But he diverged markedly from Poincaré in extending this account to mathematics and logic as well. His basis for doing so was Hilbert's formalized theory of arithmetic, which Schlick hoped (wrongly, as it turned out) would eventuate in a consistency proof for arithmetic ([1925] 1985, 357).

The conventionalism that emerged was thus one in which true propositions known *a priori* were regarded as components of autonomous symbol games that, while perhaps constructed with an eye to an application, are not themselves answerable to an independent reality (Schlick [1925] 1985, 37–8). Schlick thought that the laws of logic, such as the principles of identity, noncontradiction, and the excluded middle, "say nothing at all about the *behavior of reality*. They simply regulate how we *designate the real*" ([1925] 1985, 337). Schlick acknowledged that the negation of logical principles like noncontradiction was unthinkable, but he suggested that this fact was *itself* a convention of symbolism (concerning the notion 'unthinkable'), claiming that "anything which contradicts the principle is termed *unthinkable*" ([1925] 1985, 337).

Although Schlick thought that the structure of a symbol system, including inference systems such as logic, was autonomous and established by a set of implicit definitions, he also recognized the possibility that some of its primitive terms could be coordinated by ordinary definitions with actual objects and properties. In this way, a conventionally established symbol system could be used to describe the empirical world, and an object designated by a primitive term might be empirically discovered to have previously unknown features. Nonetheless, some of the properties and relations had by a primitive term would continue to be governed by the system conventions through which the term is implicitly defined (Schlick [1925] 1985, 48ff). These properties and relations would thus be knowable *a priori.*

The conventionalism of Schlick's *General Theory of Knowledge* anticipated many of the conventionalist elements of the position advanced by members of the Vienna Circle (see Vienna Circle). Philosophers such as Hahn, Ayer, and Carnap saw in conventionalism the possibility of acknowledging the existence of necessary truths known *a priori* while simultaneously denying such propositions any metaphysical significance. Inspired by Wittgenstein's analysis of necessary truths in the *Tractatus*, Vienna Circle members identified the truths of mathematics and logic with *tautologies*—statements void of content in virtue of the fact that they hold no matter what the facts of the world may be (Hahn 1980; Ayer 1952). Tautologies, in turn, were equated with analytic statements, and Circle members regarded all analytic statements as either vacuous conventions of symbolism known *a priori* through stipulation or as derivable consequences of such conventions knowable *a priori* through proof (see Analyticity).

An especially noteworthy outgrowth of the conventionalism of the Vienna Circle was Carnap's *The Logical Syntax of Language*. In this book Carnap advanced a conventionalist treatment of the truths of mathematics and logic through his espousal of the **principle of tolerance**, which states:

> *In logic, there are no morals.* Everyone is at liberty to build up his own logic, i.e., his own form of language, as he wishes. All that is required of him is that . . . he must state his methods clearly, and give syntactical rules instead of philosophical arguments.
>
> (Carnap 1937, 52)

Carnap regarded a language as a linguistic framework that specified logical relations of consequence among propositions. These logical relations are a precondition of description and investigation. In this view, there can be no question of justifying the selection of an ideal or even unique language with reference to facts, since any such justification (including a specification of facts) would presuppose a language. Mathematics is no exception to this. As a system of "framework truths," mathematical truths do not describe any convention-independent fact but are consequences of the decision to adopt one linguistic framework over another, in Carnap's account.

Conventionalism about mathematics and logic has faced a number of important objections. A significant early objection was given by Poincaré (1952, 166–72), who rejected Hilbert's idea that the principle of mathematical induction might merely be an implicit definition of the natural numbers (see Hilbert [1935] 1965, 193). Mathematical induction should not be regarded in this way, Poincaré argued, for it is presupposed by any demonstration of the consistency of the definition of number, since consistency proofs require a mathematical induction on the length of formulas. Treating the induction principle as itself conventional while using it to prove the consistency of the conventions of which it is a part would thus involve a *petitio.*

Carnap's *Logical Syntax of Language* sidestepped such problems by removing the demand that a system of conventions be consistent; the principle of tolerance made no such demand, and Carnap regarded the absence of consistency proofs as of limited significance (1937, 134). It appears that he would have regarded a contradictory language system to be pragmatically useless but not impossible.

Kurt Gödel, however, raised an important objection to this strategy. Gödel claimed that if mathematics is to be "merely" a system of syntactical conventions, then the conventions must be *known* not to entail the truth or falsity of any sentence involving a matter of extralinguistic empirical fact; if it does have such entailments, in Gödel's view, the truth of such sentences is not merely a syntactical, conventional matter (Gödel 1995, 339). If the system contains a contradiction, then it will imply every empirical sentence. But according to Gödel's **second incompleteness theorem**, a consistency proof required to assure the independence of the system cannot emerge from within the system itself if that system is consistent (1995, 346).

Recent commentators (Ricketts 1994; Goldfarb 1995) have argued that a response to this objection is available from within Carnap's conventionalist framework. Gödel's argument requires acceptance of an extralinguistic domain of empirical facts, which a stipulated language system may or may not imply. But it is doubtful that Carnap would have accepted such a domain, for

Carnap considered linguistic conventions, including the conventions constitutive of mathematics, as antecedent to any characterization of the facts, including facts of the empirical world. The issues involved in Gödel's objection are complex and interesting, and a thorough and satisfactory conventionalist response remains to be formulated.

Another objection to conventionalism about logic was advanced by Willard V. Quine (see Quine, Willard Van Orman). Quine objected against Carnap that treating logical laws and inference rules as conventional truths implicitly presupposed the logic that such conventions were intended to establish. Consider for instance a proposed logical convention *MP* of the form "Let all results of putting a statement for *p* and a statement for *q* in the expression 'If *p*, then *q* and *p*, then *q*' be true." In order to apply this convention to particular statements *A* and *B*, it seems that one must reason as follows: *MP* and if *MP*, then (*A* and if *A*, then *B* imply *B*); therefore, *A* and if *A*, then *B* imply *B*). But this requires that one use *modus ponens* in applying the very convention that stipulates the soundness of this inference. Quine concluded that "if logic is to proceed *mediately* from conventions, logic is needed for inferring logic from the conventions" (Quine 1966, 97). Quine at one point suggested that similar reasoning might undermine the intelligibility of the notion of a linguistic convention in general (98–9). However, after David Lewis' analysis of tacit conventions (Lewis 1969), Quine acknowledged the possibility of at least some such conventions (Quine, in Lewis 1969, xii).

Quine's original objection suggests that the conventionalist about logical truth must have an account of how something recognizable as a convention can intelligibly be established "prior to logic." As with Gödel's objection, the issues are difficult and have not yet been given a fully satisfying conventionalist account. However, there are a variety of strategies that a conventionalist might explore. One will be mentioned here. Consider an analogous case, that of rules of grammar. On one hand, one cannot specify rules of grammar without employing a language (which has its grammatical rules). Thus one cannot acquire one's first language by, say, reading its rules in a book. On the other hand, this fact does not seem to rule out the possibility that any given rules of grammar are to some extent arbitrary, and conventionally adopted. Whether the conventionalist can extend this line of thought to a defensible form of conventionalism about logic remains to be conclusively demonstrated.

AU:43 CORY JUHL AND ERIC LOOMIS

References

Ayer, Alfred J. (1952), *Language, Truth and Logic*. New York: Dover.

Carnap, Rudolph (1937), *The Logical Syntax of Language*. London: Routledge and Kegan Paul.

Duhem, Pierre ([1906] 1954), *The Aim and Structure of Physical Theory*. Translated by Philip P. Wiener. Princeton, NJ: Princeton University Press.

Eddington, A. S (1953), *Space, Time and Gravitation*. Cambridge: Cambridge University Press.

Gergonne, Joseph Diaz (1818), "Essai sur la théorie de la définition," Annales de mathématiques pures et appliquées 9:1–35.

Giedymin, Jerzy (1982), *Science and Convention: Essays on Henri Poincaré's Philosophy of Science and the Conventionalist Tradition*. Oxford: Pergamon.

Gödel, Kurt (1995), "Is Mathematics Syntax of Language?" in Solomon Feferman (ed.), *Kurt Gödel: Collected Works*, vol. 3. Oxford: Oxford University Press, 334–362.

Goldfarb, Warren (1995), "Introductory Note to *1953/9," in Solomon Feferman (ed), *Kurt Gödel: Collected Works*, vol. 3. Oxford: Oxford University Press, 324–334.

Grünbaum, Adolph (1968), *Geometry and Chronometery in Philosophical Perspective*. Minneapolis: University of Minnesota Press.

——— (1973), *Philosophical Problems of Space and Time*, 2nd ed. Dordrecht, Netherlands: Reidel.

Hahn, Hans (1980), *Empiricism, Logic, and Mathematics*. Translated by Brian McGuinness. Dordrecht, Netherlands: Reidel.

Hilbert, David ([1935] 1965), "Die Grundlegung der elementaren Zahlenlehre," in *Gesammelte Abhandlungen*, vol. 3. New York: Chelsea Publishing, 192–195.

——— ([1921] 1962), *Grundlagen der Geometrie*. Stuttgart: B. G. Teubner.

——— (1971), "Hilbert's Reply to Frege," in E. H. W. Kluge (trans., ed.), *On the Foundations of Geometry and Formal Theories of Arithmetic*. London: Yale University Press, 10–14.

Lewis, David (1969), *Convention: A Philosophical Study*. Oxford: Blackwell.

Lie, Sophus (1959), "On a Class of Geometric Transformations," in D. E. Smith (ed.), *A Source Book in Mathematics*. New York: Dover.

Malament, David (1977), "Causal Theories of Time and the Conventionality of Simultaneity," *Nous* 11: 293–300.

Norton, John (1992), "Philosophy of Space and Time," in M. H. Salmon et. al. (eds), *Introduction to the Philosophy of Science*. Indianapolis: Hackett.

Poincaré, Henri ([1905] 1952), *Science and Hypothesis*. Translated by W. J. Greestreet. New York: Dover Publications.

——— (1952), *Science and Method*. Translated by Francis Maitland. New York: Dover Publications.

Putnam, Hilary (1975a), "An examination of Grunbaum's Philosophy of Geometry," in *Mathematics, Matter and Method Philosophical Papers*, vol. 1. New York: Cambridge University Press.

——— (1975b), "Reply to Gerald Massey," in *Mind, Language and Reality, Philosophical Papers*, vol. 2. New York: Cambridge UP.

Quine, Willard V. (1966), "Truth By Convention," in *The Ways of Paradox and Other Essays*. New York: Random House, 70–99.

Ricketts, Thomas (1994), "Carnap's Principle of Tolerance, Empricism, and Conventionalism," in Peter Clark and

Bob Hale (eds.), *Reading Putnam*. Cambridge, MA: Blackwell, 176–200.
Salmon, Wesley (1977), "The Philosophical Significance of the One-Way Speed of Light," *Noûs* 11: 253–292.
Sarkar, Sahotra, and John Stachel (1999), "Did Malament Prove the Non-Conventionality of Simultaneity in the Special Theory of Relativity?" *Philosophy of Science* 66: 208–220.
Schlick, Moritz ([1925] 1985), *General Theory of Knowledge*. Translated by Albert E. Blumberg. La Salle, IL: Open Court.
Spirtes, Peter L. (1981), "Conventionalism and the Philosophy of Henri Poincaré," Ph. D. Dissertation, Department of the History and Philosophy of Science, University of Pittsburgh.
Stein, Howard (1977), "On Space-Time and Ontology: Extract from a Letter to Adolf Grunbaum," in *Foundations of Space-time Theories*, Minnesota Studies in the Philosophy of Science, vol. 8. Minneapolis: University of Minnesota Press.

See also **Analyticity; Carnap, Rudolf; Duhem Thesis; Hilbert, David; Instrumentalism; Logical Empiricism; Poincaré, Henri; Quine, Willard Van Orman; Schlick, Moritz; Space-Time; Vienna Circle**

CORROBORATION

Karl R. Popper (1959) observed that insofar as natural laws have the force of prohibitions, they cannot be adequately formalized as unrestrictedly general **material conditionals** but incorporate a modal element of natural necessity. To distinguish between possible laws and mere correlations, empirical scientists must therefore subject them to severe tests by serious attempts to refute them, where the only evidence that can count in favor of the existence of a law arises from unsuccessful attempts to refute it. He therefore insisted upon a distinction between 'confirmation' and 'corroboration' (see Popper, Karl Raimund).

To appreciate Popper's position, it is essential to consider the nature of natural laws as the objects of inquiry. When laws of nature are taken to have the logical form of unrestrictedly general material conditionals, such as $(x)(Rx \rightarrow Bx)$ for "All ravens are black," using the obvious predicate letters Rx and Bx and $\rightarrow$, as the material conditional, then they have many logically equivalent formulations, such as $(x)(\neg Bx \rightarrow \neg Rx)$, using the same notation, which stands for "Every nonblack thing is a nonraven." As Carl G. Hempel (1965) observed, if it is assumed that confirming a hypothesis requires satisfying its antecedent and then ascertaining whether or not its consequent is satisfied, then if logically equivalent hypotheses are confirmed or disconfirmed by the same evidence, a white shoe as an instance of $\neg Bx$ that is also $\neg Rx$ confirms the hypothesis "All ravens are black" (see Induction, Problem of).

Popper argued that there are no "paradoxes of falsification" parallel to the "paradoxes of confirmation." And, indeed, the only way in which even a material conditional can be falsified is by things satisfying the antecedent but not satisfying the consequent. Emphasis on falsification therefore implies that serious tests of hypotheses *presuppose satisfying their antecedents*, thereby suggesting a methodological maxim of deliberately searching for examples that should be most likely to reveal the falsity of a hypothesis if it is false, such as altering the diet or the habitat of ravens to ascertain whether that would have any effect on their color. But Popper's conception of laws as prohibition was an even more far reaching insight relative to his falsificationist methodology.

Popper's work on *natural necessity* distinguishes it from logical necessity, where the notion of projectible predicates as a pragmatic condition is displaced by the notion of dispositional predicates as a semantic condition. It reflects a conception of laws as relations that cannot be violated or changed and require no enforcement. Fetzer (1981) has pursued this approach, where lawlike sentences take the form of **subjunctive conditionals**, such as $(x)(Rx \Rightarrow Bx)$, where $\Rightarrow$ stands for the subjunctive conditional. This asserts of everything, "If it were a raven, then it would be black." The truth of this claim, which is logically contingent, depends on a difference between permanent and transient properties. It does not imply the counterpart, "If anything were nonblack, then it would be a nonraven."

Among Popper's most important contributions were his demonstration of how his falsificationist methodology could be extended to encompass

statistical laws and his propensity interpretation of probability (see Probability). Distributions of outcomes that have very low probabilities in hypotheses are regarded, by convention, as methodologically falsifying those hypotheses, tentatively and fallibilistically. Popper entertained the prospect of probabilistic measures of corroboration, but likelihood measures provide a far better fit, since universal hypotheses as infinite conjunctions have zero probability. Indeed, Popper holds that the appropriate hypothesis for scientific acceptance is the one that has the greatest content and has withstood our best attempts at its falsification, which turns out to be the least probable among the unfalsified alternatives.

The likelihood of hypothesis H, given evidence E, is simply the probability of evidence E, if hypothesis H is true. While probabilistic measures have to satisfy axioms of summation and multiplication—for example, where mutually exclusive and jointly exhaustive hypotheses must have probabilities that sum to 1—likelihood measures are consistent with arbitrarily many hypotheses of high value. This approach can incorporate the laws of likelihood advanced by Ian Hacking (1965) and a distinction between *preferability* for hypotheses with a higher likelihood on the available evidence and *acceptability* for those that are preferable when sufficient evidence becomes available. Acceptability is partially determined by relative likelihoods, even when likelihoods are low. Popper (1968, 360–91) proposed that where E describes the outcome of a new test and B our background knowledge prior to that test, *the severity of* E *as a test of* H, *relative to* B, might be measured by

$$P(E \mid H \wedge B) - P(E \mid B), \tag{1}$$

that is, as the probability of E, given H and B, minus the probability of E, given B alone. The intent of equation 1 may be more suitably captured by a formulation that employs a symmetrical—and therefore absolute—measure reflecting differences in expectations with respect to outcome distributions over sets of relevant trials, such as

$$|\, P(E \mid H) - P(E \mid B) \,|, \tag{2}$$

which ascribe degrees of nomic expectability to relative frequency data, for example. When B entails E, then $P(E|B) = 1$, and if $P(E|H) = 1$ as well, the severity of any such test is minimal, i.e., 0. When H entails E, while B entails $\neg E$, the severity of such a test is maximal, i.e., 1. The acceptance of H may require the revision of B to preserve consistency.

A plausible measure of *the degree of corroboration* C *of* H, *given* E, *relative to* B, would be

$$c(H \mid E \wedge B) = L(H \mid E)[\,|\, P(E \mid H) - P(E \mid B) \,|\,], \tag{3}$$

that is, as the product of the likelihood of H, given E, times the severity of E as a test of H, relative to B. This is a Popperian measure, but not necessarily Popper's. Popper suggests (as one possibility)

$$C(H \mid E \wedge B) = \\ P(E \mid H) - P(E \mid B)/P(E \mid H) + P(E \mid B), \tag{4}$$

which even he does not find to be entirely satisfactory and which Imre Lakatos severely criticizes (Lakatos 1968, especially 408–16). When alternative hypotheses are available, the appropriate comparative measures appear to be corroboration ratios

$$\frac{C(H_2 \mid E)}{C(H_1 \mid E)} = \frac{L(H_2 \mid E)[\,|\, P(E \mid H_2) - P(E \mid H_1) \,|\,]}{L(H_1 \mid E)[\,|\, P(E \mid H_1) - P(E \mid H_2) \,|\,]} \tag{5}$$

that reduce to the corresponding likelihood ratios of $L(H_2|E)$ divided by $L(H_1|E)$ and assume increasing significance as a function of the severity of those tests, as Fetzer (1981, 222–30) explains.

Popper also proposed the propensity interpretation of physical probabilities as probabilistic dispositions (Popper 1957 and 1959). In a revised formulation, the single-case propensity interpretation supports a theory of lawlike sentences, logically contingent subjunctive conditionals ascribing permanent dispositional properties (of varying strength) to everything possessing a reference property (Fetzer 1981 and 1993). Propensity hypotheses are testable on the basis of the frequencies they generate across sequences of trials. Long runs are infinite and short runs are finite sequences of single trials. Propensities predict frequencies but also explain them. Frequencies are evidence for the strength of propensities.

Popper (1968) promoted the conception of science as a process of conjectures and (attempted) refutations. While he rejected the conception of science as a process of inductive confirmation exemplified by the work of Hans Reichenbach (1949) and Wesley C. Salmon (1967), his commitments to deductive procedures tended to obscure the role of ampliative reasoning in his own position. Popper rejected a narrow conception of induction, according to which the basic rule of inference is "If m/n observed As are Bs, then infer that m/n As are Bs, provided a suitable number of As are tested under a wide variety of conditions." And, indeed, the rule he rejects restricts scientific hypotheses to those couched in observational

language and cannot separate *bona fide* laws from correlations.

Although it was not always clear, Popper was not thereby rejecting induction in the broad sense of ampliative reasoning. He sometimes tried to formalize his conception of corroboration using the notion of absolute (or prior) probability of the evidence E, $P(E)$, which is typically supposed to be a subjective probability. Grover Maxwell (1974) even develops Popper's approach using Bayes's theorem. However, appeals to priors are inessential to formalizations of Popper's measures (Fetzer 1981), and it would be a mistake to suppose that Popper's account of severe tests, which is a pragmatic conception, could be completely formalizable.

Indeed, Popper's notion of accepting hypotheses on the basis of severe tests, no matter how tentatively and fallibilistically, implies ampliative reasoning. Its implementation for probabilistic hypotheses thereby requires large numbers of trials over a wide variety of conditions, which parallels the narrow inductivist conception. These results, however, must be subjected to severe tests to make sure the frequencies generated are robust and stable under variable conditions.When Volkswagens were first imported into the United States, for example, they were all gray. The narrow inductivist rule of inference justified inferring that all Volkswagens were gray, a conclusion that could not withstand severe tests.

JAMES H. FETZER

References

Fetzer, James H. (1981), *Scientific Knowledge: Causation, Explanation, and Corroboration*. Dordrecht, Netherlands: D. Reidel.

——— (1993), *Philosophy of Science*. New York: Paragon House.

Hacking, Ian (1965), *Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Hempel, Carl G. (1965), *Aspects of Scientific Explanation*. New York: Free Press.

Lakatos, Imre (1968), "Changes in the Problem of Inductive Logic," in *The Problem of Inductive Logic*. Amsterdam: North-Holland, 315–417.

Maxwell, Grover (1974), "Corroboration without Demarcation," in Paul A. Schilpp (ed.), *The Philosophy of Karl R. Popper*, part 1. LaSalle, IL: Open Court, 292–321.

Popper, Karl R. (1959), *The Logic of Scientific Discovery*. New York: Harper & Row.

——— (1957), "The Propensity Interpretation of the Calculus of Probability, and the Quantum Theory," in Stephan Korner (ed.), *Observation and Interpretation in the Philosophy of Physics*. New York: Dover Publications, 65–70.

——— (1959), "The Propensity Interpretation of Probability," *British Journal for the Philosophy of Science* 10: 25–42.

——— (1968), *Conjectures and Refutations*. New York: Harper & Row.

Reichenbach, Hans (1949), *The Theory of Probability*. Berkeley and Los Angeles: University of California.

Salmon, Wesley C. (1967), *The Foundations of Scientific Inference*. Pittsburgh: University of Pittsburgh Press.

See also **Confirmation Theory; Hempel, Carl Gustav; Induction, Problem of; Popper, Karl Raimund; Verisimilitude**

COSMOLOGY

See **Anthropic Principle**

COUNTERFACTUALS

See **Causality**

Comp. by:Kandavel Date:23/8/05 Time:22:58:21 Stage:First Proof File Path://spsind002s/spare1/PRODENV/000000~2/
00FAE5~1/S00000~1/000000~4/000000~4/000008337.3D Proof by: QC by:

Author Query Form

Book Title: Philosophy of Science: An Encyclopedia
Alphabet Name:

Query Refs.	Details Required	Author's response
AU1	Whose italics?	
AU2	Should ''s'' be subscript: ''Rs''? Ditto ''Ps''?	
AU3	This is not in Refs	
AU4	Stet? I don't think I understand usage of ''nothing above''	
AU5	These symbols didn't show up	
AU6	These symbols didn't show up	
AU7	This symbol didn't show up	
AU8	Where? In what section?	
AU9	Please check for symbols and operators in these next few lines. They didn't come out	
AU10	''where Xi is''? Define Xi?	
AU11	Please provide figure legend for figure 1.	
AU12	Citation date for Cartwright?	
AU13	Stet? Where?	
AU14	OK as edited, to differentiate ''Par'' from probability ''P''?	
AU15	Please check symbols in what follows; they didn't show up	
AU16	Stet plural?	
AU17	Please clarify this sentence. Stet?	
AU18	Blue indicates Refs not cited in text	
AU19	Page range correct?	
AU20	This is not in Refs	
AU21	Blue indicates Ref not cited in text	
AU22	''a'' or ''b''?	
AU23	This is not in Refs	
AU24	Blue indicates Refs not cited in text	
AU25	'', respectively,''?	
AU26	Correct? Not '', which''?	
AU27	symbol unclear	
AU28	OK as edited?	
AU29	Cite for quotation?	
AU30	Update of this Ref per LOC.gov	
AU31	OK as edited?	
AU32	Should x's be subscript? GLOBAL	
AU33	Page range OK? I could not verify 1959, so I updated the Ref per the ''Search Inside'' feature on Amazon. com	
AU34	Correct?	
AU35	'', where f is...''?	
AU36	OK as edited?	
AU37	OK per above?	

AU38	Blue indicates refs not cited in text	
AU39	Ref updated per Amazon.com. OK?	
AU40	Please clarify. Stet?	
AU41	‘‘1991,’’ or some other cite/ref?	
AU42	Where?	
AU43	Spelling correct? Name appears elsewhere with two n’s	
AU44	Blue means that Ref wasn’t cited in text	
AU45	Antecedent of Pronoun unclear. Clarify?	