Chapter 25
Models of the Development of Scientific Theories

Gerhard Schurz

Abstract The three basic kinds of theory development are expansion, contraction and revision by empirical evidence (Sect. 25.1). Under empiricist assumptions, the history of scientific evidence can be represented by a sequence of true and cumulatively increasing evidence sets which in the limit determine the complete structure of the world (Sect. 25.2). Under these assumptions it turns out that purely universal hypotheses are falsifiable with certainty, but verifiable only in the limit, ∀-∃-hypotheses are falsifiable in the limit but not verifiable in the limit, and ∀-∃-∀-hypotheses are neither nor (Sect. 25.3). In the consequence, hypotheses with complex quantification structure can only be confirmed probabilistically (Sect. 25.4). While these results are based on "empiricist" assumptions, the revision of theories which contain theoretical concepts requires either a given partition of possible hypotheses out of which the most promising one is chosen (rational choice paradigm), or it requires steps of abductive belief revision (construction paradigm) (Sect. 25.5). Revision of scientific theories is based on a Lakatosian preference structure, following the idea that in case of a conflict between theory and data, only peripherical parts of the theory are revised, while the theory's core is saved from revision as long as possible (Sect. 25.6). Surprisingly, the revision of a false theory by true empirical evidence does not necessarily increase the theory's truthlikeness (Sect. 25.7). Moreover, increase in empirical adequacy does not necessarily indicate progress in theoretical truthlikeness; a well-known attempt to justify this inference is Putnam's no-miracles argument (Sect. 25.8).

G. Schurz (✉)
University of Düsseldorf, Düsseldorf, Germany
e-mail: schurz@phil-hhu.de

S. O. Hansson, V. F. Hendricks (eds.), *Introduction to Formal Philosophy*, Springer Undergraduate Texts in Philosophy, https://doi.org/10.1007/978-3-319-77434-3_25

25.1 Basic Notions of Theory Development

Three different kinds of theory development are expansions, contractions and revisions of theories (compare ch. 6/5 in *this volume*). These and other distinctions require *formal tools* for their reconstruction. We assume the logical framework of a 1st order language with the standard logical symbols ($\neg,\wedge,\vee,\rightarrow,\leftrightarrow,\forall,\exists, =$) and x_i ($i = 1, 2, \ldots$) for individual variables, a_i for (individual constants), P_i, R_i … for n-ary predicates. $\mathcal{L}$ is our language and $S(\mathcal{L})$ the set of $\mathcal{L}$-sentences. Small (arabic) letters $s_i \in S(\mathcal{L})$ denote arbitrary sentences, h_i hypotheses, e_i evidence statements, T_i stands for theories which are sets of sentences, E_i for sets of evidence statements, and S_i for arbitrary sets of statements. ||— denotes logical inference and Cn logical consequence, i.e. $\mathrm{Cn}(S){:=}\,\{s_i{\in}S(\mathcal{L})\text{: } S \mid\mid\!\!-\, s_i\}$ (":=" means "identity by definition"). In this framework (which is widespread in formal philosophy of science) a theory T_i is reconstructed as a consistent and logically closed set of sentences of a given set of characteristic axioms A_i of T_i, i.e. $T_i = \mathrm{Cn}(A_i)$. A_i is also called the (axiomatic) *base* of T_i.[1]

Within this formal background the (ordinary) *expansion* of a theory T by a new and T-compatible piece of information s is denoted as $T + s$ and defined as $T + s{:}= \mathrm{Cn}(T \cup \{s\})$. If $T = \mathrm{Cn}(A)$ an equivalent definition is $T + s = \mathrm{Cn}(A \cup \{s\})$. If s contradicts T ($T \mid\mid\!\!-\neg s$), then one must first contract T by $\neg s$ before one can expand T by s. The so-called *contraction* of T by s is denoted as $T \div s$ and intended to be some preferred T-subset which does no longer entail s. Different methods of defining contraction operations have been suggested (e.g. via intersections of maxichoice contractions, or via epistemic ordering or ranking functions over sentences or over corresponding possible worlds (cf. e.g. [11, 33], and ch. 6/5 *this volume*). Finally, the (ordinary) *revision* of T by a T-incompatible proposition s is denoted as $T{*}s$ and defined via the so-called Levi-identity as $T{*}s{:}= (T \div \neg s) + s$ (so-called after Isaac Levi in [10]). In other words, one revises T by a new piece of information s by first contracting T by $\neg s$ and then expanding this contraction by s. Note that for a T-compatible s, expansion coincides with revision ($T{*}s = T + s$).

The information by which scientific theories are typically expanded or revised are *evidences*, i.e. observations or measurement results; they are (typically) expressed by so-called *basic statements*. These are unnegated or negated atomic statements (also called "literals") of an assumed *empirical* (or non-theoretical) *sublanguage* $\mathcal{L}_e$ of $\mathcal{L}$ whose concepts are directly observable or measurable (i.e. $S(\mathcal{L}_e) \subseteq S(\mathcal{L})$).

[1] An alternative to the *sentential* representation of a theory T as a logically closed set of sentences is the *model-theoretic* (or "structuralist") representation of T by the set of *semantic models* or *possible worlds* which verify T (cf. e.g. [9, 17]).

Important for philosophy of science is the distinction between *AGM-contraction* after Alchourrón et al. [1] and *base-contraction* introduced by Hansson [13]. While in AGM-contraction the contraction-operation is applied to the entire set of T's consequences, whether basic or derived, in base-contraction this operation is only applied to T's axioms A, i.e. $T \div s$ is defined as $\mathrm{Cn}(A \div s)$. The difference is this. If an axiom $s \in A$ of a theory $T = \mathrm{Cn}(A)$ is removed, then in base-contraction T-consequences whose justificational support depends on axiom s have to be removed from T, too. In AGM-contraction this need not be so: one may retain consequences from T even after their premises have been removed. While base-revision is reasonable from a *foundation-oriented* viewpoint, AGM-revision makes sense if one adopts a *coherentistic* position (cf. also [33], ch. 3).

Beyond these ordinary notions of expansion and revision (be it AGM- or base-) we introduce in Sect. 25.5 the stronger notions of *abductive* expansion and revision.

25.2 Theory Development Under Empiricist Assumptions

The empiricist view of scientific theories has been defended by classical empiricists (e.g. John Locke) and early logical empiricists (e.g. [4]). This view makes two assumptions: (i) that evidences are certain, and (ii) that all non-logical concepts of scientific theories can be defined by empirical concepts, or in other words, that the scientific language contains no genuinely theoretical concepts, i.e. $\mathcal{L} = \mathcal{L}_e$. We let $E(\mathcal{L}_e)$ stand for the set of all basic (i.e. evidence) statements of $\mathcal{L}_e$. Each maximally consistent subset of $E(\mathcal{L}_e)$ (a so-called "diagram") is denoted by EW_k and represents a possible empirical world or "$\mathcal{L}_e$-world" over a given countably infinite domain of named individuals (i.e. every individual has a standard name in $\mathcal{L}_e$). We formalize the history of scientific evidences *over* a given $\mathcal{L}_e$-world EW in the form of an *evidence stream* $(e) = (e_0, e_1, \ldots)$, such that $EW = \{e_i: i \in \omega\}$ and the e_i are pairwise distinct. Given the empiricist assumptions, EW determines the *complete true* theory $T(EW)$, which is semantically given as the set of all true $\mathcal{L}_e$-statements in EW and syntactically as the uniquely determined maximally consistent and ω-complete extension of EW in $\mathcal{L}_e$.[2] Given this formal reconstruction, empiricist theory-development has the following properties:

1. The accumulation of evidences is *cumulative* because evidences are true and are never taken back. Formally, this means that the e_i are mutually compatible.
2. At every finite time point n, only a finite subset of EW is known. Therefore $T(EW)$ is never known with certainty. As Popper [28] has stressed, no purely *universal hypotheses* (e.g. $\forall xPx$) is verifiable by finitely many evidences, although

[2]A sentence set $S \subseteq S(\mathcal{L})$ is maximally consistent iff S is consistent and has no consistent proper extension in $S(\mathcal{L})$ and S is ω-complete iff whenever $\varphi[\mathrm{a}_i] \in S$ for all individual constants a_i, then $\forall x\varphi[x] \in S$ (for $\varphi[\mathrm{a}_i]$ an arbitrary $\mathcal{L}$-formula containing a_i).

it is falsifiable (namely by $\neg Pa$); and dually, no purely *existential hypothesis* (e.g. $\exists xPx$) is falsifiable by finitely many evidences, although it is verifiable (namely by Pa). Later it was discovered that quantificationally mixed hypotheses such as $\forall x\exists xRxy$ are neither verifiable nor falsifiable by finitely many evidences. It follows that even under empiricist assumptions it need not be that theory development is cumulative at the level of hypotheses: false universal theories may be held for an arbitrarily long time before they get falsified; and theories with mixed quantifiers will never get verified or falsified.

25.3 Convergence in the Limit and Formal Learning Theory

The empiricist setting of Sect. 25.2 is the major framework of formal learning theory (cf. ch. 6/6, *this volume*). Evidence streams are called *data streams* in Kelly [15]. Formal learning theorists are aware that there exist theories with theoretical terms to which this setting doesn't apply. Still their setting may be regarded as a legitimate idealization in all contexts in which one may assume an unproblematic background knowledge (e.g. concerning measurement techniques) by which one can determine the truth value of all basic statements of the language in which theories are formulated. From now on we understand "$\mathcal{L}_e$" in this extended sense. Verification and falsification *with certainty* (in the sense of Carnap [4] and Popper [28]) are defined as follows:

(1) $h \in S(\mathcal{L}_e)$ is *verifiable* (or falsifiable, respectively) *with certainty* over an $\mathcal{L}_e$-world EW iff for every evidence stream (e) over EW there exists a time point n at which e_n entails h (or entails $\neg h$, resp.).

Purely existential hypotheses are verifiable and purely universal hypothesis are falsifiable with certainty; but quantificationally mixed hypotheses are neither nor. In view of this negative results formal learning theorists suggest to use the *weaker* epistemic standard of verifiability [viz. falsifiability] in the limit. Let $SQ(\mathcal{L}_e)$ be the set of all evidence sequences over $\mathcal{L}_e$-worlds; let (e_{1-n}) denote the sequence of the first n elements of (e); and let $E_n: = \{e_i: 1 \leq i \leq n\}$ be the corresponding evidence set at time n. An assessment function for the hypotheses $h \in H$ in a given set of hypotheses H is a function α: $H \times \{(e_{1-n}): (e) \in SQ(\mathcal{L}_e), n \in \omega\} \rightarrow$ {true, false} which conjectures at every time point n of any evidence stream (e) whether h is true or false. Then:

(2) A hypothesis $h \in S(\mathcal{L}_e)$ is *verifiable* (or falsifiable, resp.) *in the limit* iff there is an assessment function such that for every $\mathcal{L}_e$-world EW which verifies (or falsifies, resp.) h and evidence stream (e) over EW there exists a time point n after which α conjectures the correct truth value of h in EW forever (i.e. $\alpha(h,(e_{1-m}))$ = true for all $m \geq n$).

One of the major results of formal learning theory is the following:

(3) An ∃-∀-hypothesis (e.g. $\exists x \forall y Rxy$) is verifiable but not falsifiable in the limit. Dually a ∀-∃-hypotheses (e.g. $\forall x \exists y Rxy$) is not verifiable but falsifiable in the limit.

In particular, a ∀-hypothesis ($\forall x Fx$) is verifiable in the limit and falsifiable with certainty (and dually for ∃-hypotheses, with "verifiable" and "falsifiable" exchanged). Let us explain the basic idea underlying result (3). Define Dom(E_n) to be the *subdomain* of those individuals which appear in E_n. Then an assessment method α for $\exists x \forall y Rxy$ can be defined as follows: conjecture "true" as long as the so far observed evidence set E_n does not falsify $\exists x \forall y Rxy$ over Dom(E_n); otherwise conjecture "false". Then if h is true in *EW* there exists an individual a_k which appears at some time $t(a_k)$ and for which $\forall y R a_k y$ is true, whence after time $t(a_k)$ α will conjecture "true" forever. However, if $\exists x \forall y Rxy$ is false in a given *EW*, then for every assessment method α one may construct a "demonic" evidence stream over *EW* such that α's conjectures don't stabilize but switch endlessly between "true" and "false" (for the detailed construction cf. [15], 51ff).

Although the method α for $\exists x \forall y Rxy$ is guaranteed to stabilize to the conjecture "true" after *some finite* time, one can never know *when* this time is reached, and hence, one can never know whether one has achieved the truth or not. Even if this intrinsic weakness of "verification in the limit" is accepted, the other bad news is that already hypotheses with three alternating quantifiers (∃-∀-∃ or ∀-∃-∀) are neither verifiable nor falsifiable in the limit. Kelly shows that these hypotheses are at least *gradually* verifiable (or falsifiable) in the limit, which is a still weaker property (whose definition is omitted here; cf. [ibid., 66ff]). However, for hypotheses with four alternating quantifiers even gradual verification (or falsification) fails.

Most contemporary philosophers of science would argue that even evidence statements may fail: perceptions may be erroneous and measurement devices may be malfunctioning. If we drop the infallibility assumption for evidences, then evidence statements in (e) may mutually contradict each other. It follows that evidence streams are no longer cumulative: $E_n \subseteq E_{n+1}$ does not always hold and we cannot define $E_n = \{e_i: 1 \leq i \leq n\}$. We rather have to construct the *evidence history* (E) as a sequence of evidence sets $(E) := (E_0, E_1, E_2, \ldots)$ that is defined by revision (or expansion) operations: $E_{n+1} = E_n * e_{n+1}$. The axiom of success, $e_n \in E_n$, is no longer mandatory for revisions over contradicting evidence statements. To retain the positive results concerning verifiability and falsifiability it is necessary to set up the following *constraint of stable error-correction:* every false evidence in (e) is corrected once-and-forever after some finite time. This implies that for every true evidence e in *EW* (the given empirical world) there exists a time n such that for all $m \geq n$, $e \in E_m$.[3]

[3]The constraint of *error-correction in the limit* would not be sufficient.

25.4 Inductive Confirmation and Convergence with Probability

The major conclusion of the previous section is this: even under empiricist idealizations and in regard to the *weak* "in-the-limit" notions of verification and falsification, the range of those hypotheses which are verifiable or falsifiable is very *restricted*. As soon as one assumes *inductive confirmation* as a legitimate justification principle, things get better. A simple qualitative definition may run as follows:

(4) A hypothesis h is inductively *confirmed* by a finite evidence set E_n iff h is not falsified by E_n over the subdomain Dom(E_n), and this the confirmation is the stronger, the greater that part of Dom(E_n) which verifies h over Dom(E_n).[4]

However, since David Hume it is known that induction is not a generally reliable inference: it may fail in worlds (or event sequences) which are *non-uniform*, i.e. in which the future differs radically from the past. What one can only show is that (under mild conditions) induction is an *optimal* strategy (cf. [35]).

Often confirmation principles are formalized in the framework of *Bayesian* (subjective) *probabilities*, i.e. rational degrees of beliefs (cf. ch. 6/8 and 6/9, *this volume*). If $P{:}S(\mathcal{L}_e) \rightarrow [0,1]$ is a Bayesian probability function over the total set of statements, then the (posterior) probability of a hypothesis h_k given evidence e is given by the famous Bayes-formula as

(5) $P(h_k|\, e) = P(e|\, h_k) \cdot P(h)/\Sigma_{1 \leq i \leq n} P(e|\, h_i) \cdot \mathrm{P}(h_i)$

where $\{h_1, \ldots, h_n\}$ is a (pragmatically given) partition of possible hypotheses that contains P_k, $P(h_i)$ is the prior probability of h_i and $P(e|h_i)$ the probability (or "likelihood") of e on the assumption that h_i. Formula (5) shows that the posterior probability of a hypothesis depends not only on its relation to the evidence (the likelihood), but also on its prior probability. Since most contemporary Bayesians agree that prior probabilities are *subjective,*[5] this dependency seems to constitute a counterargument to Bayesian confirmation theory. Bayesians counter that the dependency of the posterior probabilities of hypotheses on the priors becomes smaller and smaller the more evidences come in. Bayesians cite here *convergence* theorems ([8], 58), for example the following general convergence theorem:

(6) Gaifman and Snir [12]: Under the assumption that P is (not only finitely but) *countably additive* it holds for every possible $\mathcal{L}_e$-hypothesis h and evidence stream over a world EW with *probability* 1 that h's posterior $P(h|E_n)$ converges to the truth value of h in EW for $n \rightarrow \infty$.

Probabilistic convergence theorems of this sort are restricted in three ways: (1.) they hold only under the *empiricist* assumption that $\mathcal{L} = \mathcal{L}_e$ (Gaifman and Snir

[4] *Example:* If $E_2 = \{Raa, Rab, Rbc\}$, then $h = \exists x \forall y Rxy$ is neither falsified nor verified by E_2 over Dom(E_2) $= \{a,b,c\}$, though it is verified over $\{a,b\}$.

[5] The older view of Carnap [5] on "logically given" prior probabilities is hardly tenable; cf. Howson/Urbach ([14], 60); Earman ([8], 15).

express this by their requirement that $S(\mathcal{L}_e)$ "separates" the set of all possible worlds over $\mathcal{L}_e$); (2.) they hold only with probability 1 − whence convergence may fail in an uncountably infinite subset of the uncountably many $\mathcal{L}_e$-worlds, and (3.) they hold only under the condition of countable additivity, which involves inductive assumptions (this is demonstrated in ([15], 321ff) and [39]). Stronger convergence properties require stronger inductive assumptions (cf. [8], 108, and Schurz [41], ch. 4.7).

25.5 The Rational Choice and the Construction Paradigm of Theory Discovery: Learning Sentences and Abductive Theory Revision

In the previous sections we have studied the development of *scientific assessments* of hypotheses in the face of an evidence stream, but not the *discovery* of hypotheses. For Popperians theory discovery is a matter of psychology, not of logic − there are no rules for theory discovery. Formal learning theory, however, provides also rules for theory discovery. These rules assume that there exists a countably enumerable set H of possible (not necessarily disjoint) hypotheses in $\mathcal{L}_e$ which contains at least one true hypothesis, and an infinitely repetitive ordering $(h) = (h_0, h_1, \ldots,)$ of the hypotheses in H, i.e. every hypothesis occurs in (h) infinitely many often ([15], 224f). With such an enumeration at hand, an assessment method $\alpha(h_i,(e_{1-n}))$ can be transformed into a discovery method γ that assigns to each initial subsequence (e_{1-n}) $(n\in\omega)$ a hypothesis in H, recursively defined as follows:

(7) $\gamma((e_0)) = h_0$, and for each time n, if $\alpha(\gamma((e_{1-n}),(e_{1-n})) =$ "true", then $\gamma((e_{1-(n+1)})) = \gamma((e_{1-n}))$, and otherwise $\gamma((e_{1-(n+1)}))$: = the next hypothesis h^* in (h) behind $\gamma((e_{1-n}))$ for which $\alpha(h^*, (e_{1-(n+1)})) =$ "true".

The so-defined discovery method γ stabilizes to conjecturing the first true hypothesis in H for which α stabilizes to the assessment "true".

This discovery rule of formal learning theory is an example of the *rational choice paradigm* of theory discovery. Here one assumes that a list H which contains all possible and interesting hypotheses is given *in advance*, i.e. *before* any evidences have been received. Often this list H is assumed to form a *partition*, i.e. the hypotheses in H are pairwise incompatible and exhaustive (at least in relation to an assumed background knowledge K). Theory development consists in choosing the optimal (e.g. best confirmed) hypotheses in the face of the received evidence set E_n.

Although in some historical phases theory development proceeds according to the rational choice paradigm, this paradigm has its limitation in the fact that scientists rarely possess a list of all interesting hypotheses in advance from which they choose. Usually new interesting hypotheses are *constructed* in science from given evidence by inductive or abductive learning mechanisms and are then put to subsequent empirical tests ([34], §1; [37], §1.3). We call this view the *construction* paradigm of theory development. To fill the construction paradigm with content we need heuristics and/or algorithms which tell us how to construct and revise plausible theories in the face of an ongoing evidence stream.

Neither the AGM- nor the base-version of ordinary belief revision contains such learning mechanisms, although on different reasons. AGM-revision is extremely liberal: it allows T^*e to be any logically closed sentence set which lies between Cn($\{e\}$) and some maximally consistent extension of $\{e\}$ and which is "preferred"; but the AGM-axioms for preferences don't decide which one of these sets is preferred. Belief base revision, on the other hand, is purely *corrective* in the sense that here one cannot generate *new* quantified hypotheses from evidences (for details cf. [37], §1.2). In scientific theory development, however, the revision process is typically *creative* in the sense that it constructs new hypotheses.

In line with the two paradigms there have been suggested two ways of extending the theory of belief (base) revision in order to cover theory discovery. One way is based on the rational choice paradigm; it has been introduced by Levi (1980, 35f; 1991, 71ff, 146) and is further developed by Rott [33] and Olsson and Westlund (2006). Besides input-driven (or "routine") revisions, Levi introduces a second kind of revisions, so-called *deliberate* revisions, which result from an act of will and consist in choosing a hypothesis h from a given partition of possible hypotheses and adding it to the accepted beliefs.

The second way is based on the construction paradigm and consists in enriching the accepted beliefs by creative learning mechanisms. This idea has been implemented in two ways, by learning sentences and by abductive revision operations. The method of *learning sentences* has been introduced by Martin and Osherson (1998) (the name "learning sentences" comes from me). Assume a cumulative sequence of evidence sets $(E) = (E_0, E_1, \ldots)$. Let $(T)\!: = (T_0, T_1, \ldots)$ be an *associated* sequence of theories (you may also say "belief system" instead of "theory") which by definition arises from (E) and an initial theory T_0 by *revision* operations: $T_n\!: = T_{n-1}{}^*E_n$. Note that we revise with E_n (instead with only the new evidences in E_n) because (a) iterated revision is not always order-independent and (b) this definition works also if (E) is not cumulative. Assume the underlying empirical world EW makes the universal hypotheses $\forall xFx$ true. The problem is that ordinary belief base revision can never generate $\forall xFx$ in the face of the evidence stream (E) as long as T_0 is empty or contains only singular statements (proof in [37], §1.2). Martin and Osherson overcome this problem by adding a learning sentence of the form $F\mathrm{a}_i \rightarrow \forall xFx$ to the initial theory T_0 (for an arbitrary constant a_i). If $\forall xFx$ is true, $\forall xFx$ will enter the theory sequence at the first time n at which $F\mathrm{a}_i$ occurs in the evidence stream, and will remain there forever, while if $\forall xFx$ is false, then $\neg\forall xFx$ will enter the theory sequence at the first time n at which a sentence of the form $\neg F\mathrm{a}_j$ has shown up in the evidence stream.

In the case of $\exists$-$\forall$-hypotheses learning sentence are more complicated than simple implications from evidences to hypotheses. More generally Martin and Osherson prove the following *theorem*:

(8) [Martin and Osherson 1998, (63), p. 153] For each problem of the form "which of the hypotheses in partition $\{h_1, \ldots, h_n\}$ is true in *EW*" that is solvable by a discovery algorithm γ (in the explained sense) there exists a set of learning sentences L such stringent belief base revision[6] applied to a cumulative evidence sequence (E) over *EW* and an initial belief set T_0 containing $\{\vee_{1 \leq i \leq n} h_i\} \cup L$ will after some finite time n produce belief sets T_n, $T_{n+1}, \ldots$ which contain the *true* element h^* of $\{h_1, \ldots, h_n\}$ forever (i.e., $h^* \in T_m$ for all $m \geq n$).

Martin and Osherson develop a fascinating combination of formal learning theory and belief revision. However, their account has two problems. First, it is restricted to the empiricist assumption $\mathcal{L}_e = \mathcal{L}$. Second, learning sentences are somehow unnatural: we do not literally believe "if this (and this …) raven is black, then all ravens are black".

An alternative way of implementing learning mechanism is by closing the revision operation under non-deductive inferences. If these non-deductive inferences include abductions to conclusions with theoretical concepts, hypothesis creation transcends the empiricist assumption. Abductive belief revision has been elaborated, among others, by Pagnucco [27], Aliseda [2], Schurz [37] (see also Langley et al. [20]). Abductive belief expansion can be defined as follows (cf. [37], §3.2):

(9) Let $T(\mathcal{L})$ be the set of all possible theories in the given total language $\mathcal{L}$ (which may now contain theoretical terms; i.e. $S(\mathcal{L}_e) \subset S(\mathcal{L})$). The *abductive expansion* of a theory $T \in T(\mathcal{L})$ by a T-compatible evidence e is denoted by $T +_a e$ and defined as follows: $(T +_a e) = (T + e) + \text{abd}(T,e)$, where "+" is the ordinary expansion function.

In this context, $abd(T,e)$: $T(\mathcal{L}) \times E(\mathcal{L}_e) \rightarrow S(\mathcal{L})$ is an abductive expansion function that assigns to each theory T and evidence e a consistent sentence $s := abd(T,E)$ which is either a tautology or *explains* E within K.

This definition allows for the case in which no explanatory hypothesis for E is found; in that case $abd(T,e)$ is identified with a tautology (and the abductive expansion is called "improper").

Abductive expansion can be broken up into an ordinary expansion and an abductive inference step (this is in line with what ([33], 84) calls the "direct mode of foundationalist base revisions"). The same is not always possible for abductive revision. If an element h of belief set T explains a conjunction of evidences E, and e is a new piece of evidence contradicting h, then it is inefficient to remove h and generate an alternative hypothesis h^* from scratch. Scientists rather try to obtain the revised hypothesis h^* by a *direct* revision of the old hypothesis h given $E \cup \{e\}$ into some new hypothesis h^* which explains e and at the same time preserves the old explanations of the evidences in E. For example, assume h is a quantitative hypothesis saying that gas pressure is proportional to the gas temperature, and new

[6] A base contraction function $\div$ is stringent iff for each T and e, $T \div e$ is a preferred maximal T-subset not implying e.

data tell us that for low temperatures, the gas pressure is lower than predicted by h. Then scientists will not simply remove h from their belief set, but revise h by adding a new non-linear term to the linear relationship predicted by h. This leads to the following general definition of abductive belief revision "$*_a$":

(10) $T*_a\, e := T*e + \text{rev}(h,e, T\div\neg e)$ (where "*" is the ordinary revision operator).

Here $h := h_{T,e}$ is the "explanatory loss" caused by the contraction by e, defined as the conjunction of all explanatory hypotheses which are in T but not in $T\div\neg e$. Moreover, rev: $T(\mathcal{L}) \times E(\mathcal{L}_e) \times \text{Pow}(S(\mathcal{L})) \rightarrow S(\mathcal{L})$ is an abductive revision function which assigns to the "lost" hypothesis h, h-incompatible evidence e and theory-contraction $T\div\neg e$ a revised hypothesis $h^* := \text{rev}(h,e, T\div\neg e)$ that is consistent with $T*e$ and is either a tautology or explains $E_{T,h}\cup\{e\}$, where $E_{T,h}$ is the set of all evidences which were explained by h in T.

In the process of hypothesis-revision, the revised hypothesis h^* is not only a function of the contracted theory $T\div\neg e$ and the new evidence e – which it would have to be according to Levi identity – but also a function of the old hypothesis h which has been removed from $T\div\neg e$. Schurz ([37], §3.2) concludes from this fact that Levi-identity fails for abductive belief revision.

25.6 Theoretical Concepts, Lakatosian Research Programs and Refined Falsificationism

Most scientific theories contain *theoretical concepts* such "electron" or "magnetic force" which do not occur in the evidence stream but go beyond the observable. For theoretical hypotheses – i.e. hypotheses containing theoretical terms – the empiricist assumption fails: their truth value is not determined by the evidence stream, even not in the limit; moreover, they are not obtainable from evidences by inductive generalizations from finite evidence sets. How are theoretical hypotheses confirmed at all?

Popper [28] has pointed out that usually scientific theories entail observational consequences and are, though not being verifiable, at least falsifiable via the rule of *Modus tollens* (if $T \mid\mid\!\!- E$ then $\neg E \mid\mid\!\!- \neg T$). Popper's falsificationist account of theory development rests on the idea that theories which are falsified by the actual evidences are laid aside. Popper's account of "instantaneous rejection" was criticized by Kuhn [16] and Lakatos [19]. Kuhn showed that in the history of science theories which contradict data are not rejected or laid aside; scientists rather introduce additional *auxiliary assumptions* which save the theory core or theory "paradigm" from being falsified. It is well known that Kuhn's criticism involved some more radical points, for example concerning the theory-dependence of evidence and the resulting irrationality of paradigm changes (so-called "scientific revolutions"). Many (and probably most) contemporary philosophers of science did not follow these radical aspects of Kuhn.

The less radical part of Kuhn's criticism was elaborated by Lakatos' account of *refined* falsificationism which significantly improved Popper's "naive account". Scientific hypotheses are never assessed in isolation. They form theories, which are *systems* of statements together with an *epistemic ranking* (preference ordering) over them. This ordering decides which elements are to be given up when conflicts between theory and data arise. Lakatos speaks of theories as consisting of a *theory core* that is surrounded by a *periphery* which contains less and less important parts, the more one moves from inside into outside layers of the theory. The outermost layer of a theory contains auxiliary assumptions, which assert the existence or non-existence of *disturbing factors*. They figure like a *protective belt* because by introducing new disturbing factors the theory core can always be protected from falsification. For example, in 1846, when J. Adams and U. Le Verrier discovered a considerable discrepancy between the predicted and the actually observed orbit of the planet Uranus, they postulated the existence of a yet undiscovered planet, Neptune, whose gravitational effect on Uranus was assumed to be responsible for its divergence from the predicted orbital path. Later on Neptune's existence was indeed independently confirmed by telescopic observations. But a similar scenario happened around 1856 when Le Verrier observed a divergence of the planet Mercury from its predicted orbit and postulated the existence of a yet smaller planet named *Vulcan*, which despite tenacious attempts could never be found by telescopes.

Similar accounts of theory development had already been given by Duhem [7] and Quine [32]. Duhem's thesis of the *holism of theory falsification* says that if a particular *version* T of a theory – consisting of T's core plus a particular periphery – contradicts a datum e, then all what one knows (by Modus Tollens) is that some part of T is false, but logic alone doesn't dictate which part of T should be given up. Lakatos [19], however, provided an answer to this question: the theory-parts which are given up should be as peripherical and unimportant as possible (this is a version of "prioritized base contraction" in the sense of Rott ([33], 40ff). Although it is logically speaking always possible to solve conflicts with data ("anomalies") by peripherical theory-revisions, Lakatos sets up an important rationality constraint to steps of this sort: a theory-revision should be *theoretically progressive*, by which Lakatos means that the new theory contains all the confirmed empirical content of the old theory plus some additional new empirical "excess" content. Moreover, Lakatos calls the new theory version *empirically progressive* if part of this excess content has been independently confirmed. If, on the other hand, theory revisions *reduce* the empirical content of a theory they are called *degenerative*.

In Lakatos' account of research programmes a theory is a *historical* entity: if it changes its periphery, it is merely another *version* of the same theory; only if it changes its core, it is a *new* theory. Formally speaking, a *theory history* in the sense of Lakatos is a sequence $(T) = (T_0, T_1, \ldots)$ which is associated with a sequence of evidence sets $(E) = (E_0, E_1, \ldots)$ in the sense defined in the pervious section. Each theory $T_i = \mathrm{Cn}(A_i)$ is now itself a ranked system of statements ($T_i(0)$, $T_i(1)$, $T_i(2), \ldots$), where $T_i(0) \subset A_i$ is the *core* of theory T_i and $T_i(n)$ ($n \geq 1$) are less and less important subsets of T-axioms. Building on the previous section we assume that each T_n is given as the abductive revision of the background theory T_0 by E_n ($T_\mathrm{n}{:}= T_0{*}E_\mathrm{n}$).

In line with Popper and Lakatos we do not assume that evidence statements are infallible, i.e. the evidence sequence (*E*) is not necessarily cumulative. Of course we make the Lakatosian assumption that the evidence sequence – though not being absolutely theory-neutral – is at least *theory-neutral* in regard to all those theories whose success is being compared. Only under that assumption can one have the *same* evidence sequence for all theories in (*T*), even for those with a different core. This is the decisive difference to Kuhn's relativistic account in which different theories with different cores would have *their own* evidence sequences, and rational comparisons of them were hardly possible. On the other hand, the decisive difference of Lakatos' account to neo-empiricist accounts is that the theories in (*T*) contain theoretical concepts, whence their truth is not determined by an empirical world *EW* over $\mathcal{L}_e$.

Theory-subsequences of (*T*) whose theories share the same theory core are called (in line with Kuhn) *normal* periods of science; while theory-successions (or pairs) in which the theory core changes are called scientific *revolutions*. We define the following Lakatosian criteria for the rational evaluation of theory-development (where "$\subset$" stands for proper and "$\subseteq$" for proper-or-improper set-inclusion):

(11) We define $E(T) = T \cap S(\mathcal{L}_e)$ as the set of all (confirmed or unconfirmed) empirical consequences of *T*, and $EC_n(T) := E(T) \cap E_n$ as the set of *T*'s confirmed empirical consequences at time *n*.

(12) A theory succession (T_n, T_{n+1}) is called

- *theoretically progressive* iff $E(T_n) \subset E(T_{n+1})$ and $EC(T_n) \subseteq EC(T_{n+1})$;
- *empirically progressive* iff $E(T_n) \subset E(T_{n+1})$ and $EC(T_n) \subset EC(T_{n+1})$;
- *stagnating* iff $E(T_n) = E(T_{n+1})$ and $EC(T_n) = EC(T_{n+1})$, and finally
- *degenerative* iff either $E(T_n) \supset E(T_{n+1})$ or $EC(T_n) \supset EC(T_{n+1})$.

In conclusion, Lakatos' model of theory development allows for the rational evaluation of theory development and the rational assessment of theory progress even if one allows for fallible evidences and for theories whose truth value is not determined by the complete empirical truth.

25.7 Verisimilitude and Truth-Approximation by Theory-Revision

Popper (1963, 233f) has argued that the main goal of science consist in progress in *verisimilitude* or *truthlikeness*. For although most scientific theories involve idealizations and hence are strictly speaking false, some of them are much *closer* to the truth than others. According to Popper's intuitive idea, a theory T_1 is closer to the truth than another theory T_2 iff T_1 has more true and less false consequences than T_2. It is well known that Popper's original definition of verisimilitude turned out to be inadequate (cf. ch. 7/2, *this volume*). In the following period two major families of accounts to verisimilitude have been developed which cured the mistake in Popper's original definition; they have been called the *disjunctive* and the *conjunctive* approach to truthlikeness (cf. [6, 18, 40]). Without being able to explain the precise definitions of verisimilitude in these accounts, we mentioned some major results concerning the connection between verisimilitude and belief revision:

(12) Niiniluoto [25]: Neither the expansion nor the revision of a false theory T by a true evidence leads always to an increase of T's verisimilitude.

Schurz [38] demonstrates that this holds even if the evidence is a single true basic statement (instead of a disjunction of a true and a false basic statement, as assumed in Niiniluoto's example). The following example makes this clear: assume the hypothesis $h := b \rightarrow f_1 \wedge \ldots \wedge f_n$ is an implication leading from a true but unknown basic statement b to a conjunction of false basic statements, and the new input by which the theory $T := \{h\}$ is expanded is b. Then $T + b = \mathrm{Cn}(\{b, f_1, \ldots, f_n\})$. Since the verisimilitude-loss due to the addition of n false basic statements f_i $(1 \leq i \leq n)$ may be much greater than the verisimilitude-gain due to the addition of the true b, the verisimilitude of $T + b$ may be much smaller than that of T. However, the volume Kuipers and Schurz [38] contains a lot of results which show that under restricted conditions positive connections between increase in verisimilitude and revision by true evidences can be obtained. In conclusion, truth-approximation can still be upheld as a major goal of theory development, although the paths towards this goal may have intermittent phases in which theories move away from the truth.

25.8 From Progress in Empirical Success to Progress in Theoretical Truth: Instrumentalism Versus Realism

Scientific realism is the view that the empirical success of a theory is a reliable indicator of the (approximate) truth of the theory, including the truth of its theoretical superstructure. In contrast, *scientific instrumentalism* holds that the theoretical superstructure of a theory has merely the instrumental purpose of entailing the evidences in a most simple and unifying way, but there is no reason to assume that this theoretical superstructure corresponds to an unobservable external reality. While for scientific realists the decisive progress in theory development consists in progress in truth approximation at the theoretical level, for instrumenalists or empiricists such as [42] scientific progress is confined to progress in *empirical success* (or empirical adequacy).

The standard justification of scientific realism is the *no-miracles* argument, which goes back to Putnam (1975, 73) and has been used in various ways as a defense of scientific realism (cf. Boyd 1984; [30]). This argument says that the empirical success of contemporary scientific theories would be a sheer miracle if we would not assume that their theoretical superstructure, or ontology, is approximately true in the sense of scientific realism. However, the no-miracles argument is beset by two strong counterarguments, an empirical and a theoretical counterargument.

The empirical counterargument is the *pessimistic meta-induction* argument of Laudan (1981). This argument points to the fact that in the history of scientific theories one can recognize radical changes at the level of theoretical superstructures, although there was continuous progress at the level of empirical success. On simple inductive grounds one should expect therefore that the theoretical superstructures of our presently accepted theories will also be overthrown in the future, and hence can in no way be expected to be approximately true.

The theoretical counterargument to the no-miracles argument is the *no-speculation* argument (cf. [34], §7.1). It points out that for every set of possible observations E one may construct *ex-post* and *ad-hoc* some speculative theory T which just entails ("explains") E, but has no independent empirical consequences. The empirical success of such speculative ad-hoc theories is in no way a reliable indicator of their approximate truth.

In Schurz [36] a justification of the inference from empirical success to partial theoretical truth has been suggested which does not presuppose the questionable NMA. It is based on relations of *correspondence* between historically consecutive theories T and T^* with nondecreasing empirical success, which hold if the following (simplified) conditions are satisfied:

(C1): The predecessor theory T speaks about a partition of circumstances $\{A_1, \ldots, A_n\}$ and contains a theoretical expression φ for which it entails a set of bilateral reduction sentences $\{B_i{:}1 \leq i \leq n\}$ of the form

(B_i): $\forall x \forall t(A_i xt \rightarrow (\varphi(x) \leftrightarrow R_i xt))$ – in words: for all systems x and times t, under empirical circumstances A_i the presence of φ in system x is indicated by the empirical phenomenon R_i.

(C2): Every empirical prediction of the form $\exists t(A_i xt \wedge R_i xt) \rightarrow \forall t(A_j xt \rightarrow R_j xt)$ which follows from $\{B_i{:}1 \leq i \leq n\}$ is entailed by the successor theory T^* in a *T^*-dependent* way, which means by definition that there exists a theoretical *mediator* description $\varphi^*_{i,j}x$ in T^* such that $\forall x \exists t(A_i xt \wedge R_i xt \rightarrow \varphi^*_{i,j}x)$ as well as $\forall x(\varphi^*_{i,j}x \rightarrow \forall t(A_j xt \rightarrow R_j xt))$ follows from T^*.

In addition it is required that the two theories T and T^* are *causally normal* in the sense that the circumstances A_i are described in terms of theory-independent empirical parameters. While condition (C2) is mild, condition (C1) on the predecessor theory T is a crucial constraint which requires that the theoretical concept φ of T figures as a common cause of several empirical regularities. The collection of the bilateral reduction sentences $\{B_i{:}1 \leq i \leq n\}$ entailed by T is denoted as B(T,φ). (C1) excludes pre-scientific ad-hoc speculations from the application range of the correspondence theorem. Given these conditions it is possible to infer that also a part of the theoretical structure of T is preserved in T^*:

(13) *Correspondence theorem* [36]: Let T be a causally normal predecessor theory satisfying condition (C1) and T^* a causally normal successor theory satisfying condition (C2). *Then* T^* contains a theoretical expression $\varphi^*(x)$ such that $B(T,\varphi) \cup T^*$ is consistent and implies a *correspondence relation* of the form $\forall x \forall t(A_1 xt \vee \ldots \vee A_n xt \rightarrow (\varphi(x) \leftrightarrow \varphi^*(x)))$. In words: whenever a system x

is exposed to one of the circumstances A_i, then x satisfies the T-theoretical description φ iff x satisfies the T*-theoretical description φ*.

Based on the correspondence theorem (13) one can argue that φ(x) *refers indirectly* to the theoretical state of affairs described by φ*(x) − provided one assumes that φ*(x) refers and T* is at least partially true. An example which is extensively discussed in Schurz [36] is the transition from the *phlogiston* theory to the modern *generalized oxidation* theory of combustion and saltification. For this theory-transition the correspondence theorem generates the following correspondence: *phlogistication* of a substance x corresponds to the acceptance of electrons by positively charged x-ions from their bonding partner, and *dephlogistication* of x corresponds to the *donation of electrons* of x's atoms to their electronegative bonding partner.

The correspondence theorem allows that T and T* are mutually incompatible. However, that part of T which is needed to derive the correspondence, namely B(T,φ) is always compatible with T*. If one assumes that the observed phenomena are caused by an external reality whose structure can possibly be represented by an ideal but unknown "super-theory" T^+ which is causally normal and satisfies condition (C2), then (13) implies that also our presently accepted theories, as long as they are causally normal and satisfy condition (C1), must have got something right at their theoretical level. In conclusion, the correspondence theorem justifies a weak kind of scientific realism which is not based on the questionable no-miracles argument.

References[7]

1. Alchourrón, C. E., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change. *Journal of Symbolic Logic, 50*, 510–530.
2. Aliseda, A. (2006). *Abductive reasoning*. Dordrecht: Springer.
4. Boyd, R. (1984). The current status of scientific realism. In J. Leplin (Ed.), *Scientific realism* (pp. 41–82). Berkeley: University of California Press.
4. Carnap, R. (1928/2003). *The logical structure of the world and pseudoproblems in philosophy*. Chicago: Open Court.
5. Carnap, R. (1950). *Logical foundations of probability*. Chicago: University of Chicago.
6. Cevolani, G., & Festa, R. (2009). Scientific change, belief dynamics and truth approximation. *La Nuova Critica, 51*(52), 27–59.
7. Duhem, P. (1908/1991). *The aim and structure of physical theory*. Princeton: Princeton University Press.
8. Earman, J. (1992). *Bayes or bust?* Cambridge, MA: MIT Press.
9. French, S. (2008). The structure of theories. In S. Psillos & M. Curd (Eds.), *The Routledge Companion to Philosophy of Science* (pp. 269–280). London: Routledge.
10. Gärdenfors, P. (1981). An epistemic approach to conditionals. *American Philosophical Quarterly, 18*, 203–211.

[7] Asterisks (*) indicate recommended readings.

11. Gärdenfors, P. (1988). *Knowledge in flux*. Cambridge, MA: MIT Press.
12. Gaifman, H., und Snir, M. (1982). Probabilities over rich languages. *Journal of Symbolic Logic, 47*, 495–548.
13. * Hansson, S. O. (1999). *A textbook of belief dynamics*. Dordrecht: Kluwer. [Introduction in contemporary accounts of belief revision]
14. * Howson, C., und Urbach, P. (1996). *Scientific reasoning: The bayesian approach* (2nd ed.). Chicago: Open Court. [Introduction to probability theory with a focus on Bayesianism]
15. * Kelly, K. T. (1996). *The logic of reliable inquiry*. New York: Oxford University Press. [Textbook in formal learning theory]
16. Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press (3rd edition 1996).
17. Kuipers, T. A. F. (2000). *From instrumentalism to constructive realism*. Dordrecht: Kluwer.
18. Kuipers, T. A. F. & Schurz, G. (2011). Belief revision aiming at truth approximation. *Erkenntnis*, *75*, 2011. (Guest-edited volume).
19. Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). London: Cambridge University Press.
20. Langley, P., et al. (1987). *Scientific discovery. computational explorations of the creative process*. Cambridge, MA: MIT Press.
21. Laudan, L. (1981). *A confutation of convergent realism* (pp 107–138). Reprinted in D. Papineau (Ed.), (1996). *The philosophy of science*. Oxford: Oxford University Press.
22. Levi, I. (1980). *The enterprise of knowledge*. Cambridge, MA: MIT Press.
23. Levi, I. (1991). *The fixation of belief and its undoing*. Cambridge, MA: Cambridge University Press.
24. Martin, E., & Osherson, D. (1998). *Elements of scientific inquiry*. Cambrigdge, MA: MIT Press.
25. Niiniluoto, I. (1999). Belief revision and truthlikeness. In B. Hansson et al. (Eds.), *Internet Festschrift for Peter Gärdenfors.* http://www.lucs.lu.se/spinning
26. Olsson, E. J., & Westlund, D. (2006). On the role of research agenda in epistemic change. *Erkenntnis*, 65, 165–183.
27. Pagnucco, M. (1996). *The role of abductive reasoning within the process of belief revision* (Dissertation). Sydney: University of Syndey.
28. Popper, S. K. (1935/2002). *Logic of discovery*. London: Routledge, 2002.
29. Popper, K. (1963). *Conjectures and refutations*. London: Routledge.
30. * Psillos, S. (1999. *Scientific realism. How science tracks truth.* London/New York: Routledge. [Comprehensive overview of the contemporary debate in scientific realism]
31. Putnam, H. (1975). What is mathematical truth? In H. Putnam (Ed.), *Mathematics, matter and method* (pp. 60–78). Cambridge: Cambridge University Press.
32. Quine, W. v. O. (1951). Two dogmas of empiricism. *Philosophical Review, 60*, 20–43.
33. Rott, H. (2001). *Change, choice and inference: A study in belief revision and nonmonotonic reasoning*. Oxford: Clarendon Press.
34. Schurz, G. (2008a). Patterns of abduction. *Synthese, 164*(2008), 201–234.
35. Schurz, G. (2008b). The Meta-Inductivist's winning strategy in the prediction game: A new approach to Hume's problem. *Philosophy of Science, 75*, 278–305.
36. Schurz, G. (2009). When empirical success implies theoretical reference: A structural correspondence theorem. *British Journal for the Philosophy of Science, 60*(1), 101–133.
37. Schurz, G. (2011a). Abductive belief revision in science. In E. Olsson & S. Enqvist (Eds.), *Belief revision meets philosophy of science* (pp. 77–104). New York: Springer.
38. Schurz, G. (2011b). Verisimilitude and belief revision. With a focus on the relevant element account. *Erkenntnis, 75*(2011), 203–221.

39. Schurz, G., & Leitgeb, H. (2008). Finitistic and frequentistic approximations of probability measures with or without sigma-additivity. *Studia Logica, 89*(2), 258–283.
40. Schurz, G., & Weingartner, P. (2010). Zwart and Franssen's impossibility theorem holds for possible-world-accounts but not for consequence-accounts to verisimilitude. *Synthese, 172*, 415–436.
41. * Schurz, G. (2013). *Philosophy of science: a unified approach*. New York: Routledge. [Cromprehensive overview of contemporary philosophy of science, divided into introductory and advanced parts]
42. * Van Fraassen, B. (1980). *The scientific image.* Oxford: Clarendon Press. [Representative work of contemporary empiricism]