Ilkka Niiniluoto

BELIEF REVISION AND TRUTHLIKENESS

1. *Introduction*

In December 1977, a conference on ”The Logic and Epistemology of Scientific Change” was organized in Helsinki (see Niiniluoto and Tuomela, 1979). Employing tools from logic, set theory, and probability theory, the participants developed three different approaches to scientific change. One group (including myself) tried to see how the concept of truthlikeness could be explicated without stumbling on the difficulties of Popper´s original definition. Another group, led by Wolfgang Stegmüller, advocated the structuralist ”non-statement view” of scientific theories. The third group (among them Isaac Levi and Peter Gärdenfors) studied changes of probabilistic belief systems.

In the next decade, these research programmes produced works like Ilkka Niiniluoto´s *Truthlikeness* (1987) and Peter Gärdenfors´s *Knowledge in Flux* (1988). It is my impression that these books have been widely read by different audiences. Indeed, their approaches seems to be very far from each other, as the former relies heavily on the notion of truth, while the latter eschews this concept altogether.

However, in this paper I attempt to make some comparisons between the research programmes of truthlikeness and belief revision. A natural common ground for such a comparison is that both of them can be viewed as accounts of theory change in science. The definition of truthlikeness leads to a dynamic theory of scientific progress (see Niiniluoto, 1984). On the other hand, the principles of rational belief revision can be assessed by making explicit the hidden assumptions about the role of truth as one of the goals of inquiry.

We shall also see that the concept of similarity, which has been a key ingredient of the theory of verisimilitude, has been employed to give a semantic account of belief revision. The similarities between these approaches have not been exploited, however.

I restrict my attention here to cases which can be represented without probability functions. In *Knowledge in Flux*, this covers the treatment of expansions, revisions, and contractions (Ch. 3) and epistemic entrenchment (Ch. 4). In *Truthlikeness*, this covers the logical definition in terms of truth values and similarity, but not the estimation of truthlikeness relative to available evidence. The interesting comparison between the revision of probabilistic epistemic states and the expected degrees of truthlikeness has to be left to another paper.

2. *Belief in Flux*

According to Gärdenfors (1988), pp. 18-19, rational belief systems to a large extent mirror or represent an external reality. However, his strategy is to separate such "objective" factors from the "internal" ones dealing with the relations between belief and language. From within a belief system there is no difference between knowing that something is true and fully believing it. Therefore, Gärdenfors argues, the concepts of truth and falsity become irrelevant for the analysis of belief change.

The elegant AMG theory of belief revision supports the thesis that "many epistemological problems can be attacked without using the notions of truth and falsity" (p. 20).

Still, one might make at least the terminological complaint that the traditional notions of *episteme* and *knowledge* (from Plato to Hintikka) have distinguished knowledge from justified belief by the requirement of truth. This strong notion of knowledge might be replaced with a weaker concept where ´truthlikeness´ or ´approximate truth´ plays the role of strict ´truth´ (see Niiniluoto, forthcoming). But a philosopher should see the danger of confusion in the trend, today popular in

Artificial Intelligence and the sociology of science, to use the terms ´knowledge´ and ´belief˝ as equivalent.

Let K be a *belief set*, represented by a consistent and deductively closed set of sentences within a given language L. Hence, Cn (K) = K, and K is what is usually called a "theory" in L. A sentence A is accepted if it belongs to K, rejected if its negation ¬A belongs to K, and indetermined otherwise.

Let A be an epistemic *input* sentence, expressing in L information that we learn from some source. Assume that A is consistent with the initial belief set K. Then the *expansion* K^{+}_{A} of K by A is defined by

(1) $K^{+}_{A} = Cn (K \cup \{A\})$.

If A contradicts K, then some beliefs in K have to be revised to maintain consistency, assuming that A will be included in the new belief set. The *revision* K^{*}_{A} of K by A is then characterized by means of a "minimal" change of K to accomodate for A. The *contraction* K^{-}_{A} of K with respect to A in K is obtained by retracting from K the sentence A (with sentences which entail A). According to the *Levi identity*, revision can be defined in terms of contraction and expansion:

(2) $K^{*}_{A} = (K^{-}_{\neg A})^{+}_{A}$.

According to the *Harper identity*, contractions can be defined by revisions:

(3) $K^{-}_{A} = K \cap K^{*}_{\neg A}$.

One way of defining revisions and contractions, proposed by Adam Grove (1988) (see also Gärdenfors, 1988, Ch. 4.5), is to employ David Lewis´s techique of "spheres of similarity". For Lewis, this is a method of expressing similarities between possible worlds, so that his spheres are centered on a single possible world. Grove construes such spheres so that they are centered on a given belief set K. For this purpose, it is useful to identify possible worlds with maximally consistent sets of sentences in language L, and to represent propositions as sets of possible worlds.

Let A be a proposition, and S_A the smallest sphere that intersects A. Let $C_K(A) = A \cap S_A$ be the "closest" elements of A to K (understood also as a set of possible worlds). Grove´s idea is to define the revision K^*_A of K by A as the theory generated by $C_K(A)$, i.e., K^*_A is the intersection of all the maximally consistent sets of sentences in $C_K(A)$. This is equivalent to the condition

$B \in K^*_A$ iff $S_{A\&B} \subseteq S_{A\&\neg B}$,

i.e., B belongs to the revision K^*_A iff A&B is "closer" to K than A&¬B.

As Gärdefors shows, if the ordering of "epistemic entrenchment" is defined by

$B \leq A$ iff $S_{\neg B} \subseteq S_{\neg A}$,

then in the contraction $K^-_{A\&B}$ of K by A&B the less entrenched of A and B is given up. Thus,

$B \leq A$ iff $B \notin K^-_{A\&B}$,

This is equivalent to defining the contraction K^-_A as the theory generated by $K \cup C_K(A)$.

3. *Truthlikeness*

The first formulation of the similarity account of truthlikeness by Risto Hilpinen in 1974 employed Lewis´s spheres (see Hilpinen, 1976). My suggestion was to replace possible worlds with

constituents of a finite first-order language, and to define explicitly the *distance* Δ between such constituents.

In the simplest case, L is a language with primitive propositions $p_1, p_2, \dots, p_n$, and the constituents have the form $(\pm)p_1 \,\&\, (\pm)p_2 \,\&\, \dots \,\&\, (\pm)p_n$, where each p_i occurs as unnegated or negated. The distance between such constituents is simply the number of their differences in the $(\pm)$-signs, divided by n. In a monadic first-order logic with Q-predicates $Q_1, \dots, Q_m$, the constituents have the form $(\pm)(\exists x)Q_1(x) \,\&\, \dots \,\&\, (\pm)(\exists x)Q_m(x)$.

The constituents $C_1, C_2, \dots, C_k$ of language L are mutually exclusive and jointly exhaustive. The consequence class $Cn(C_i)$ of a constituent is a complete theory in L, i.e., a maximally consistent set of sentences of L. Each sentence in L can be expressed in a normal form as a finite disjunction of constituents: if A is a sentence in L, and **A** is the index set of the L-constituents which entail A, then

(4) $A \equiv \bigvee_{i \in \mathbf{A}} C_i$.

Hence

$$A\&B \equiv \bigvee_{i \in \mathbf{A} \cap \mathbf{B}} C_i$$

$$A \vee B \equiv \bigvee_{i \in \mathbf{A} \cup \mathbf{B}} C_i$$

$$\neg A \equiv \bigvee_{i \notin \mathbf{A}} C_i.$$

Following Hintikka, these normal forms can be extended to full first-order logic (relative to a fixed quantificational depth). The hardest part of the logical theory of truthlikeness has been to define the distance between depth-d constituents.

Let Δ_{ij} be the distance between constituents C_i and C_j. Thus, $0 \leq \Delta_{ij} \leq 1$, and $\Delta_{ij} = 0$ iff i=j. For a given constituent C_i, the other constituents C_j will be located around C_i at the distances Δ_{ij}. This defines a quantitative version of the Lewis-type system of spheres of similarity.

Let C_* be the *true* constituent of L. Then the distances Δ_{*i} tell how close to the truth the constituents C_i are. For a statement A in L, the degree of *approximate truth* of A depends on the minimum distance from A´s normal form to the target C_*:

(5) $\Delta_{\min}(C_*,A) = \min_{i \in \mathbf{A}} \Delta_{*i.}$

The degree of *truthlikeness* of A combines the idea of closeness to the truth with the demand that a good theory should exclude "bad" possibilities which are far from the truth. It can be defined in terms of the weighted average of the minimum distance from C_* and the normalized sum of all distances from C_{*i}:

(6) $\gamma\Delta_{\min}(C_*,A) + \gamma' \sum_{i \in \mathbf{A}} \Delta_{*i} / \sum_{i=1}^{k} \Delta_{*i},$

where $0 < \gamma < 1$, $0 < \gamma' < 1$.

Definition (6) yields also a comparative criterion for one theory A to be *more truthlike* than another theory B (relative to target C_*). This criterion can be applied as a definition of *cognitive progress* in science.

We are now ready to translate the principles of belief revision into our framework.

4. *Expansion*

Let T be a theory in language L, and let A be an input sentence in L, where both can be expressed in the normal form (4). Assume that A is compatible with T, so that $\mathbf{T} \cap \mathbf{A} \neq \varnothing$. Then the expansion of T by A is, according to (1), simply the conjunction of T and A:

(7) $T_A^+ =_{df} T \,\&\, A$.

Hence, T_A^+ entails T, and $T_A^+ = T$ if A is logically true or T entails A. In a genuine expansion, T&A is logically stronger than T. If A entails T, then $T_A^+ = A$.

The nature of expansion with respect to truth values can be summarized by the following principle:

(8) T_A^+ is true iff both T and A are true.

Thus, all expansions of a false theory T are false, and the expansion of a true theory T by a false input A is false. The only "safe" case is the one where a true input is added to a true theory.

A finer analysis of the expansion process can be made with the concept of truthlikeness. Note first that principles of the form

(9) If T and A are truthlike, then T_A^+ is truthlike

are not generally valid, already for the reason that two mutually incompatible sentences can both be truthlike; the same holds for approximate truth. For the same reason, the expansion of a true theory T by an input A which is false but truthlike need not be truthlike.

An important feature of our definition (6) is that among false theories truthlikeness does not covary with logical strength. As Pavel Tich_ and Graham Oddie observed in their "child´s play argument", otherwise it would be too easy to improve a false theory: just add to it an arbitrary false sentence like ´The moon is made of cheese´. Many attempts to rescue Popper´s definition of truthlikeness have failed to block this undesirable feature. In terms of expansions, our definition leads to the result that T_A^+ need not be (but may be) more truthlike than T when T and A are false – this depends on the location of $\mathbf{T} \cap \mathbf{A}$ with respect to the true target C_*.

Suppose that our false theory T states that the number of planets is 10 or 20, and let A be the true input sentence that this number is 9 or 20. Then T&A states that the number of planets is 20. This is clearly less truthlike than T. (Examples of this sort have been proposed also by Dr. Ilkka Kieseppä.) Hence we see that

(10) If T is false and A is true, T_A^+ may be less truthlike than T.

On the other hand, our definition (6) satisfies the principle that among true theories truthlikeness covaries with logical strength. Thus,

(11) If T and A are true, then T_A^+ is at least as truthlike as T.

This principle is not satisfied by the average measure advocated by Tich_ and Oddie.

5. *Revision and Contraction*.

Grove´s semantic for theory change can be explicated in our framework by employing distances between constituents. Let T be the target theory. Then the distance of a constituent C_i from T can be defined by

(12) $\Delta_i(T) = \min_{j \in \mathbf{T}} \Delta_{ij} = \Delta_{min}(C_i,T)$

(cf. (5)). Hence, $\Delta_i(T) = 0$ if $i \in \mathbf{T}$. The constituents C_i with $i \notin \mathbf{T}$ are located in spheres centered on the set of constituents in the normal form of T.

The elements of A closest to T, i.e., the set $C_T(A)$, can now be defined directly by

(13) $C_T(A) = \{i \in \mathbf{A} \mid \Delta_i(T) \leq \Delta_j(T) \text{ for all } j \in \mathbf{A}\}$.

The revision T^*_A of theory T by A is then

(14) $T^*_A =_{df} \bigvee_{i \in C_T(A)} C_i$.

As $C_T(A) \subseteq \mathbf{A}$, the revision T^*_A of T by A entails A. It also follows that $C_T(A)$ is a singleton if **A** is. In other words, revision of a theory T by a constituent C_i is identical to C_i.

Note that definition (14) includes expansion (7) as a special case: if A is compatible with T, then $C_T(A) = \mathbf{T} \cap \mathbf{A}$ and $T^*_A = T\&A = T^+_A$.

For example, let L contain three primitive propositions p, q, r, and let T be the theory p&q. Suppose that A is ¬q&r, so that T and A are incompatible. In this case T is the disjunction of

constituents p&q&r and p&q&¬r, and the set $C_T(A)$ contains the constituent p&¬q&r which is at distance 1/3 from T. Hence, the revision of p&q by ¬q&r is p&¬q&r. If T is q, then $C_T(A)$ contains both p&¬q&r and ¬p&¬q&r, so that the revision of q by ¬q&r is ¬q&r.

If A is entailed by T, the contraction T^-_A of T with respect to A can be defined by

(15) $T^-_A =_{df} \bigvee_{i \in \mathbf{T} \cup C_T(\neg A)} C_i = T \vee T^*_{\neg A}$.

Hence, the contraction T^-_A is entailed both by the original theory T and by the revision $T^*_{\downarrow A}$. Here (15) entails that

$Cn(T^-_A) = Cn(T) \cap Cn(T^*_{\neg A})$,

so that (15) corresponds directly to the Harper identity (3). The Levi identity (2) holds when $\mathbf{T} \cap \mathbf{A} = \varnothing$, since

$(T^-_{\neg A})^+_A = (T \vee T^*_A) \,\&\, A = \bigvee_{i \in [\mathbf{T} \cup C_T(A)\ ] \cap \mathbf{A}} C_i = \bigvee_{i \in C_T(A)} C_i = T^*_A$.

For example, let L again be the language with propositions p, q, r, and let T = p&q and A = q. Then $C_T(\neg A)$ includes constituents p&¬q&r and p&¬q&¬r, and T^-_A is (p&q) ∨ (p&¬q) or simply p. Thus, the contraction of p&q with respect to q is p. Similarly, the contraction of p&q with respect to p→q is p.

According to (14), revision by false information always leads to a false theory:

(16) If A is false, then T^*_A is false.

Even a revision by a true sentence A may lead from a false theory to a false theory, since the true constituent C_* may fail to belong to $C_T(A)$. However, in the special case where A is the maximally informative truth C_*, we know that T^*_A is C_* and, hence, true.

The connection of revision (and the relation of epistemic entrenchment) to the truthlikeness ordering depends upon where the true constituent C_* is located relative to **T** and **A**. Therefore, there are no general results which would guarantee that revision even by a true sentence increases truthlikeness:

(17) If T is false and A is true, T^*_A may be less truthlike than T.

To illustrate this possibility, let T state that the number of planets is 19, and A that it is 9 or 20. Then A is true, but $C_T(A)$ claims that the number of planets is 20.

According to (15), the contraction of a true theory leads to a logically weaker true theory. Thus,

(18) If T is true, then T^-_A is true.

(19) If T is true, then T is at least as truthlike as T^-_A.

On the other hand, contracting a false theory T with respect to a false sentence A in T may (but need not always) lead to a true theory T^-_A. Even though T^-_A is logically weaker than T, sometimes it may be more truthlike than T. This will always be the case if $C_T(\neg A)$ is a singleton consisting of C_*. More generally,

(20) If T is false and $C_T(\neg A) = \{i\}$, where $\Delta_{*i} < \Delta_{min}(C_*,T)$, then T^-_A is more truthlike than T.

6. *Conclusion*

Two main conclusions emerge from the previous sections.

First, using the similarity metrics employed in the theory of truthlikeness, the semantic characterization of belief revision in principle covers full first-order logic. Details of such an approach remain to be worked out.

Secondly, expansion and revision do not always lead our false beliefs closer to the truth, even when the new information is true. But they may increase the truthlikeness of our theory. To specify conditions, which make cognitive progress within belief revision at least plausible, should be a fundamental problem of epistemology.

Department of Philosophy

P.O. Box 24

00014 University of Helsinki

Finland

ilkka.niiniluoto@helsinki.fi

BIBLIOGRAPHY

Grove, A., ´Two Modellings for Theory Change´, *Journal of Philosophical Logic* 17 (1988), 157-170.

Gärdenfors, P., *Knowledge in Flux: Modeling the Dynamics of Epistemic States*, The MIT Press, Cambridge, MA, 1988.

Hilpinen, R., ´Approximate Truth and Truthlikeness´, in M. Przelecki, K. Szaniawski, and R. Wojcicki (eds.), *Formal Methods in the Methodology of the Empirical Sciences*, D Reidel, Dordrecht, 1976, pp. 19-42.

Niiniluoto, I., *Is Science Progressive?*, D. Reidel, Dordrecht, 1984.

Niiniluoto, I., *Truthlikeness*, D. Reidel, Dordrecht, 1987.

Niiniluoto, I., ´Verisimilitude: The Third Period´, *The British Journal for the Philosophy of Science* 49 (1998), 1-29.

Niiniluoto, I., *Critical Scientific Realism*, Oxford University Press, Oxford, forthcoming.

Niiniluoto, I. and Tuomela, R. (eds.), *The Logic and Epistemology of Scientific Change*, Acta Philosophica Fennica 30: 2-4, North-Holland, Amsterdam, 1979.