Logique & Analyse **133-134** (1991), 23-39

MORE SUBTLE THEORY CHANGE

Frank DÖRING

In this paper, I derive a syntactic procedure for revising theories in propositional logic from considerations of indifference and informational economy (minimality). The procedure is very flexible. It allows us to make use of information about the relative epistemic merit (entrenchment) of the sentences in the theories whenever such information is available, and, unlike other procedures proposed in the literature, yields plausible results even for very simple entrenchment orderings.(1)

1. *The problem of theory change*

The problem of theory change I propose to consider is twofold: how can old information be deleted from a given theory (a data base, stock of beliefs, knowledge state, or what have you), and how can surprising news be added? The first aspect of the problem concerns *contraction*, the second *revision*. The problem arises for every fallible representational system, hence for every —artificial or natural— agent. A representational system, even one that never commits a logical mistake, becomes fallible by drawing inferences that take it beyond the evidence. It therefore needs a method to accommodate unexpected news and give up erroneous information without giving up too much else. (Error correction is not the same as recovery from contradiction. I will not offer a recipe for resolving contradictions, I will offer only a recipe to help avoiding them.)

For present purposes, theories are consistent, logically closed sets of sentences in the propositional language L. L has a finite stock of sentence letters A, B, etc. and contains as connectives only negation and disjunction. (The restriction to the finite case is for the sake of simplicity; I shall indicate where an infinite language would call for a more complex treatment.) Sentences or *clauses* in L are sets of literals interpreted as disjunctions, and

(1) Research for this paper was supported by grants from the Fyssen foundation and the CNRS. Thanks to Marshall Farrier, Gary Gates, and François Lévy for comments on an earlier draft.

literals are sentence letters with or without negation. $\{A, \neg B\}$ for example is the disjunction $(A \vee \neg B)$. Sets of sentences in standard propositional calculus notation can be translated effectively into into L.(2) Occasionally, I will make use of the standard notation.

Logics, classical as well as non-monotonic, are of little help in our problem. A logic gives interesting directives about how to *expand* sets of sentences, namely by closing them under some relation of implication. This is relevant when we want to accommodate new information that is consistent with the theory at hand. But our problem is not with expansion, it is with contraction and revision. It is a problem about giving up information, and no logic tells us anthing about that. Logics, conceived as theories of implication, furnish state laws for objects in theory space, but they do not determine trajectories through the space.

We will write T_A^+ for the expansion of theory T by A, T_A^- for the contraction of T by A, and T_A^* for the revision of T by A. Revision can be conceived of as a composite operation: first, T is made consistent with A, *i.e.* contracted by $\neg A$; then $T_{\neg A}^-$ is expanded by A into $T_{\neg A}^- {}_A^+ = T_A^*$. This definition of revision in terms of contraction and expansion has come to be known as the Levi identity. Alternatively, contraction can be defined in terms of revision by means of the Harper identity $T_A^- = T \cap T_{\neg A}^*$. The two identities are equivalent given the Alchourrón-Gärdenfors-Makinson postulates for all three operations.(3) I find the Levi identity intuitively more compelling and will therefore treat contraction for the most part as basic and revision as composite.

Theory change should be minimal. This tenet has been labelled variously *principle of informational economy*, *conservatism*, or *minimality*. The idea is that information should be neither adopted, nor given up, gratuitously. To violate the principle is to sabotage the project of finding out about the world. This is why. Information is not self-authenticating (except in degenerate cases). Thus, if one were to adopt any of it gratuitously, one would have no reason to believe it to be true, that is no reason to believe it, that is no reason to adopt it. Information is not self-refuting either (except in degenerate cases). Thus, if one were to give up any of it gratuitously, there would be no point in gathering it in the first place. Conservatism is built into the very foundations of epistemic rationality.

(2) A translation procedure is described in Genesereth and Nilsson (1987).

(3) See for instance Gärdenfors (1988); the notation is adopted from this book.

The principle determines expansion: $T \overset{+}{_A}$ should contain the entire content of T and A, and nothing else. So $T \overset{+}{_A} = Cn\ (T \cup \{A\})$. Contraction, by contrast, is only constrained, but not determined. A semantic illustration shows why:

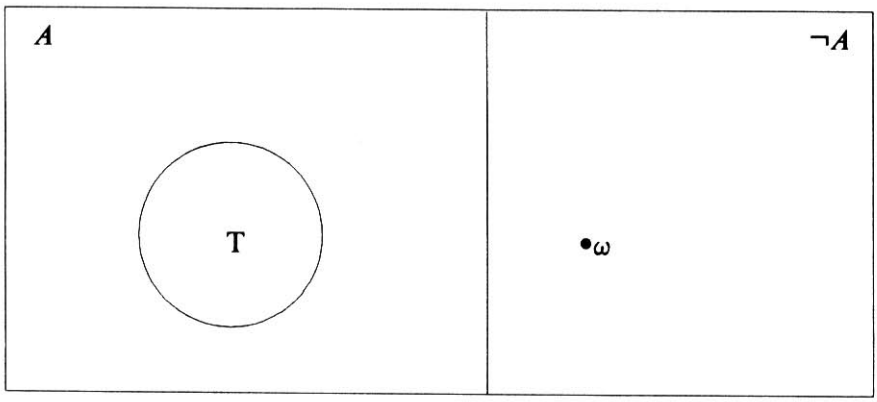

Figure 1

Think of figure 1 as depicting the set of all "possible worlds" describable in L. Contents correspond to sets of worlds or points, *i.e.*, to areas in the diagram. (We say that a clause is *valid* in a set of worlds iff it is true in every world in that set; a set of worlds W *characterizes* a theory T iff all clauses in T are valid in W and all clauses valid in W are in T, written as $W = |T|$.) Contracting theory T by A in a most conservative way means enlarging $|T|$ by exactly one point from $|\neg A|$, say ω. $|T| + \omega$ then characterizes $T \overset{-}{_A}$. If we now expand $T \overset{-}{_A}$ by $\neg A$, *i.e.* , shrink $|T \overset{-}{_A}|$ by $|A|$, we are left with ω only. The result of a most minimal contraction by A followed by an expansion by $\neg A$, which is the same as a revision by $\neg A$, is thus a maximally specific theory. We have created information out of nothing and thus violated the principle of informational economy. Radically conservative contraction has the embarrassing consequence that every revision creates information in this way.

ω was chosen arbitrarily. *A priori*, all points in $|\neg A|$ are on a par; there is nothing in T that distinguishes between them. So the alternative to selecting one arbitrary point is to select all points. This choice sets $T \overset{-}{_A}$ to $|T| + |\neg A|$, and $T \overset{*}{_{\neg A}}$ to $|\neg A|$ alone. In words: the revision of T by $\neg A$ contains nothing but $\neg A$'s logical consequences. This way, information is thrown away gratuitously. Less radical contraction is therefore as bad as its

radical alternative. For a better solution, we have to break the symmetry between the points.

A common move at this point is to deploy an ordering of *epistemic entrenchment* over the sentences in L that induces a partition of the area surrounding |T| in the model. The sentences in the representation language are assumed to be partially or totally ordered under a relation of entrenchment that mirrors their relative epistemic merits. This ordering is then used to define contraction and revision functions that retract the sentences in the order of their entrenchment, the least entrenched first and the most entrenched last.([4]) Let $\leq_T$ be an ordering relation defined over all clauses in L that satisfies the following four conditions:([5])

(1≤) For all A, B, and C, if $A \leq_T B$ and $B \leq_T C$, then $A \leq_T C$ (transitivity).
(2≤) For all A and B, if A implies B, then $A \leq_T B$ (dominance).
(3≤) For consistent T, $A \notin T$ iff $A \leq_T B$ for all B (bottom).
(4≤) If $A \leq_T B$ for all A, then B is a theorem (top).

The ordering is relative to T because according to (3≤) the sentences with the lowest ranking are just those that are not contained in T. $\leq_T$ induces a system of nested spheres around T with entrenchment increasing from the center to the periphery (see figure 2).

([4]) *E.g.* Grove (1988), Gärdenfors & Makinson (1988), Rott (1991). The appeal to something like entrenchment may be implicit. For instance, Papini's (1992) procedure works roughly by computing maximal subsets of the original theory that do not imply the sentence that is to be retracted. The procedure gives automatically more credit to sentences that are explicitly represented than to those that are merely implied, which amounts to treating explicit sentences as more entrenched. When more than one maximal subset is found, the choice among them is left to the user. I find this move unsatisfactory.

([5]) These are four of Gärdenfors's (1988) five conditions. For a discussion, see *ibid.*, pp. 89 ff.

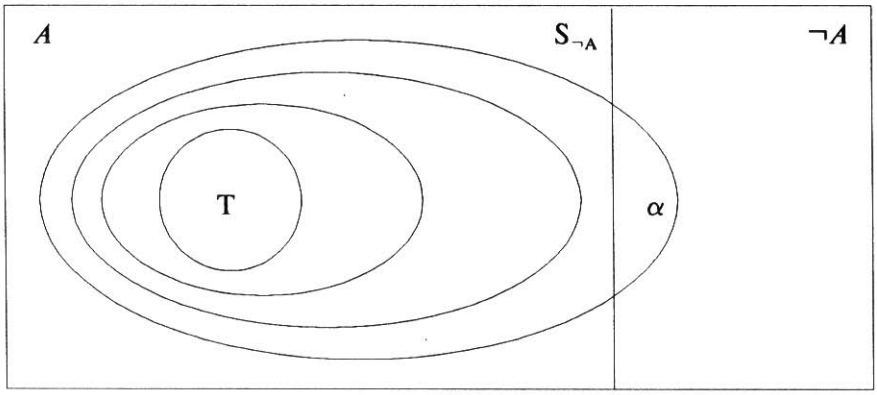

Figure 2

Each sphere S_i in figure 2 corresponds to a subset of sentences in T. The bigger the sphere, the smaller the set of sentences it characterizes and the more entrenched its least entrenched elements. Formally, if S_i and S_j are spheres such that $S_i \subset S_j$, there is at least one sentence x valid in S_i such that $x <_T y$ for every sentence y valid in S_j. The innermost sphere is $|T|$ itself, the outermost sphere (the box of figure 2) is the set of all tautologies in L. The minimal contraction of T by A can now be defined as the union of $|T|$ and the intersection α of $|\neg A|$ and $S_{\neg A}$, the smallest sphere intersecting $|\neg A|$. The revision by A is just α. Since L is finite, there is always going to be one such smallest sphere for contingent A.(6) The values of the contraction and revision function for any sentence A can be read off the system of spheres by noting which sets of sentences the spheres characterize. The set characterized by $S_{\neg A}$ consists of exactly those sentences that are strictly more entrenched than A. Revision therefore becomes: $T^*_{\neg A} = Cn(\{\{\neg A\}\} \cup \{x | x >_T A\})$, and contraction can be defined using the Harper identity $T^-_A = T \cap T^*_{\neg A}$.

In this approach to revision, it is the ordering that does all the work. The plausibility of the derived revision and contraction functions depends crucially on the details of the sphere system induced by the ordering. We therefore need some principled account of how to construct that ordering. Reference

(6) The infinite case is trickier. One may decide to impose what Lewis (1973) calls the *limit assumption* in order to assure the existence of a smallest $|\neg A|$-intersecting sphere.

to a notion of epistemic entrenchment is one attempt to fill this bill. On a natural construal of entrenchment, however, this is not going to work. In Gärdenfors's view, from which I take my bearings, it should be "possible to determine the relative epistemic entrenchment of the sentences [...] *independently* of what happens [...] in contractions and revisions" (Gärdenfors 1988, p. 87). This much seems uncontroversial for any substantial notion of epistemic entrenchment. Gärdenfors also claims, somewhat more controversially, that the entrenchment of a sentence is connected with "how useful it is in inquiry and deliberation" (*ibid.*). I have two complaints. First, an ordering of sentences in terms of their usefulness in inquiry and deliberation will normally not determine a plausible revision function *via* the sphere construction just described. Second, more generally, any ordering that has been determined independently of considerations about revision will yield a poor revision function *via* the same construction.

As for the first complaint, suppose there are two sentences A and B in T that intuitively have nothing to do with each other; say, A is about French cuisine and B about life an mars. Because of their independence, it is reasonable to demand that $A \in \mathrm{T}^{-}_{B}$ and $B \in \mathrm{T}^{-}_{A}$. The demand is satisfied just in case there is a sphere bigger than $|\mathrm{T}|$ in which the disjunction of A and B is valid but neither A nor B is, *i.e.* iff $\{A, B\}$ is strictly more entrenched than A and B individually. (*Proof.* If $\{A, B\}$ is valid in $S_{\neg A}$, then $\{A, B\} \in \mathrm{T}^{*}_{\neg A}$ and, by closure, $B \in \mathrm{T}^{*}_{\neg A}$. Since B is in T, it is also in $\mathrm{T} \cap \mathrm{T}^{*}_{\neg A} = \mathrm{T}^{-}_{A}$. On the other hand, if $\{A, B\}$ is not valid in $S_{\neg A}$, then $S_{\neg A}$ contains at least one $(\neg A \,\&\, \neg B)$-world, which lies in α. Hence B is not valid in α and so not contained in T^{-}_{A}. The reasoning for A is the same.) But surely the disjunction of A and B is no more useful in inquiry or deliberation —and thus hardly more entrenched— than A or B individually. It is difficult to imagine any other independent specification of entrenchment on which the disjunction would be strictly more entrenched than its disjuncts. In order to make things come out nicely, the ordering apparently has to be set up with an eye to its use in revision, which is what we are told to avoid.

As for the second complaint, there are certain simple entrenchment orderings that should not be ruled out *a priori*. But these orderings yield unreasonable contraction and revision functions. For example, suppose that the sentences in L are arranged in just three tiers, tautologies at the top; contingent sentences in T in the middle, and the remainder of L at the bottom. The system of spheres induced by this ordering yields the second of the two "bad" revision functions discussed before where $\mathrm{T}^{*}_{\neg A} = Cn\,\{\{\neg A\}\}$ for contingent $A \in \mathrm{T}$. The result is less drastic for a richer ordering, but the

basic flaw is the same: too much information is deleted. Since I agree with Gärdenfors that entrenchment should be kept independent of considerations about revision, I have to reject the proposed construction.

Entrenchment information would be useful nonetheless if we had a way to select a sensible *subset* of the worlds in the intersection of $|\neg A|$ and $S_{\neg A}$, that is, if we could distinguish sensibly between sentences undistinguished by the entrenchment relation. This is of course just our original problem, only now confined to a part of $|\neg A|$. In the sequel, I will develop a solution to the general problem. The solution can then be combined with a sphere model to define plausible revision and contraction functions.

2. *The model theoretic clue*

Our problem is to draw a distinction among the $\neg A$-worlds in figure 1 based on no other information than what is contained in T. It turns out that the language of T provides the constraint we need if we think of the worlds in the model in terms of their descriptions in L rather than as unstructured points. "Possible worlds" describable in L are given by maximal, consistent sets of atomic sentences and their negations. (A set is maximal just in case it contains every atomic sentence of the language in affirmed or denied form.) A theory in L is characterized again by a set of possible worlds, as depicted in Figure 3:

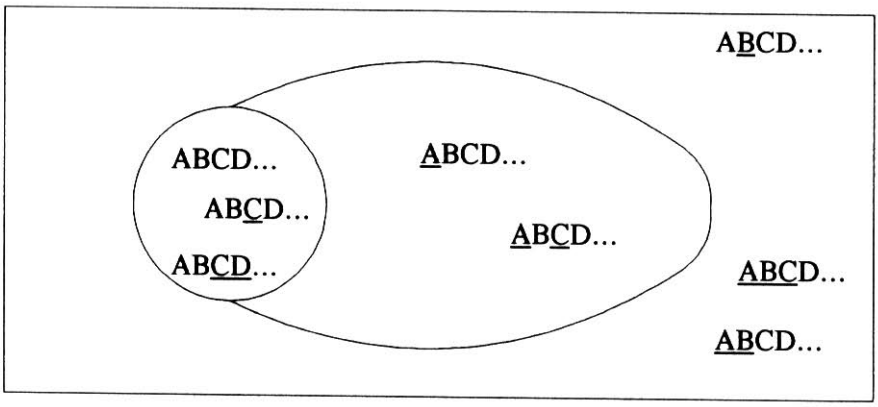

Figure 3

The items in figure 3 stand for classes of worlds, not for individual worlds (which can be very complex). For example "A$\underline{\text{B}}\underline{\text{C}}$D..." stands for the class of worlds in which A and $\neg B$ and $\neg C$ and D are true (underlining indicates negation). Theory T corresponding to the encircled area contains A and B together with their logical consequences. The $\neg A$-worlds outside $|\text{T}|$ are no longer on a par: there is a subset $|\text{T}^{\sim}|$ of them, inside the elliptic region, whose members are "closer" to the worlds in $|\text{T}|$ than any other $\neg A$-worlds. They are closer in the sense that each of $|\text{T}^{\sim}|$'s members differs from some world in $|\text{T}|$ *only* in $\neg A$. The structure of the language thus singles out a set of $\neg A$-worlds closest to T:

(C1) A non-empty set $\text{S}^{\sim}$ of $\neg P$-worlds is *closest* to a set of P-worlds S just in case for every member $s^{\sim}$ of $\text{S}^{\sim}$ there is some member s of S with which $s^{\sim}$ agrees in all respects except those relevant for the truth of P. (If P is atomic, s and $s^{\sim}$ differ just in P; if P is a disjunction, s and $s^{\sim}$ differ in one or more of P's disjuncts.)

The information contained in T and A provides no grounds for finer distinctions; we are indifferent *vis à vis* possible further subdivisions within the set of closest $\neg A$-worlds. In contracting T by A in a minimal way, we are therefore led to add the set of closest $\neg A$-worlds as a whole to $|\text{T}|$. Every smaller inclusion would impose an arbitrary partition on $|\neg A|$, and every larger inclusion would violate minimality. The resulting theory $\text{T}_{\overline{A}}$, comprising the area enclosed by the circle and the ellipsis, shows neither of the earlier defects: the theory $\text{T}_{\overline{\neg A}}^{*}$ gained by restricting $\text{T}_{\overline{A}}$'s area to its $\neg A$-worlds is neither maximally informative, nor does it consist only of the consequences of $\neg A$. $\text{T}_{\overline{\neg A}}^{*}$, corresponding to the area outside the circle and inside the ellipsis, contains all and only the logical consequences of $\neg A$ and B, which is intuitively correct.

Contraction by a disjunction works in just the same way. The selection of closest worlds falsifying a given disjunction is unique according to (C1) because there is only one way for a disjunction to be false, *i.e.* one common feature of the selected worlds: all disjuncts must be false there. The contraction rule for arbitrary clauses in L can therefore be stated as follows:

Con $\text{T}_{\overline{C}}$ is characterized by the union of $|\text{T}|$ and the set of $\neg C$-worlds closest to T.

The rule is stated in model theoretic terms. It would be better to have a

statement in syntactic terms that tells us how to revise the sentential representations of theories. Our next task is therefore to translate *Con* into a syntactic rule. This rule has to effect the removal of all clauses from T that are not valid in the set of closest $\neg C$-worlds $W^{\sim}$.

Let $C = \{L_1, \ldots, L_n\}$ be the clause in T that is to be retracted. $W = |T|$, and $W^{\sim}$ is the closest set of worlds in all of whose elements C is false, *i.e.* all the L_i are false. Which clauses can remain in T, and which have to be removed? Consider first those clauses whose intersection with C is empty, that is clauses containing none of the L_i. By assumption, these clauses are valid in W. They are also valid in $W^{\sim}$ because by (C1) every world in $W^{\sim}$ agrees with some world in W on everything except the L_i, and this agreement is all that matters for the truth of L_i-free clauses. The clauses must therefore be included in $T_{\overline{C}}$. Second, consider the clauses whose intersection with C is not empty, and among them those that contain the denial of at least one of the L_i. These clauses are valid in $W^{\sim}$ by virtue of their $\neg L_i$ disjuncts and must therefore be retained as well. This leaves us with those clauses whose intersection with C is not empty and which do not contain the denial of at least one of the L_i. These shall be all withdrawn. This move may seem too sweeping, but we will see presently that it is not. The rule is:

R1 ($\{L_1, \ldots, L_n\}$, T) If T does not imply $C = \{L_1, \ldots, L_n\}$, do nothing. Otherwise remove all clauses from T that contain any subset of C but not the denial of any of the L_i.

Atomic clauses are a limiting case of disjunctions. An atomic clause $\{A\}$ is retracted simply by removing all clauses from T that contain A as a disjunct.

R1 leaves only clauses in the remainder of T that are valid in *Con* (*C*, W) $= W \cup W^{\sim}$. Therefore all clauses in the closure of *R1* (*C*, T) must be valid in *Con* (*C*, W). If we can also show the converse, we have proved the following equivalence result for *Con* and *R1* plus closure:

Equ If a set of worlds W characterizes a theory T, then *Con* (*C*, W) characterizes *Cn* (*R1* (*C*, T)).

Assuming that W characterizes T, we have to show that every clause D that is valid in *Con* (*C*, W) is included in *Cn* (*R1* (*C*, T)). We will first establish that a clause D valid in W is not valid in *Con* (*C*, W) if its L_i-free part $D - C$ is not valid in W. Suppose that $D - C$ is not valid in W, *i.e.* false in

some world $w \in W$. There exists another world w' that is just like w except that it falsifies all the $L_i \in C$. In w', $D - C$ must be false as well, and so must be D because the literals by which it exceeds $D - C$ are all false in w'. But by (C1) w' is in $W^{\sim}$, so that D is not valid in $W^{\sim}$ and hence not valid in *Con* (C, W). The rest is straightforward. Suppose that D is valid in *Con* (C, W). Then, by contraposition of what we just proved, $D - C$ is valid in W, hence contained in T, not affected by *R1*, and consequently included in *R1* (C, T). $D - C$ implies D, and therefore $D \in Cn$ (*R1* (C, T)).

Logical closure restores what *R1* removes without need. *Equ* allows us to define contraction as follows:

$T \bar{}_C =_{def} Cn$ (*R1* (C, T)), for clauses C.

3. *Conjunctions*

A complete account of revision requires an additional rule for sets of clauses interpreted as conjunctions. Conjunctions are not sentences of L, but we need a rule for retracting what is expressed by a conjunction in other languages because otherwise we could not guarantee consistency for expansion by a disjunction: before we can expand a theory by $\{\neg C_1, \ldots, \neg C_n\}$, we have to make sure that it does not imply all of the C_i. Contraction by sets read as conjunctions raises two specific problems. One is that there is more than one way for a conjunction to be false, which opens several possibilities for interpreting the notion of minimal change; the other is that contraction and revision procedures ought to treat logically equivalent sentences alike. Let us start with the second problem. We will adopt the following postulate (read '... $\Leftrightarrow$ ...' as '...is logically equivalent to...'):

If $A \Leftrightarrow B$, then $T \bar{}_A = T \bar{}_B$.(7)

The postulate is met trivially for *R1* because every non-tautological clause in L is logically equivalent to no other clause than itself. Equivalent sets (conjunctions) of clauses, however, need not be identical, as for example

$$\{\{A, B\}, \{A, \neg B\}\} \Leftrightarrow \{\{A\}, \{A, C\}\} \Leftrightarrow \{\{A\}\}.$$

(7) For a defense, see *e. g.* Gärdenfors (1988).

(Equivalent *theories* are identical because they are closed under implication.) Contraction must be made indifferent to the different ways of expressing the same proposition. This is best achieved by rewriting the sets of clauses in a unique, canonical form before submitting them to the syntactically sensitive contraction prodedure. We define the *canonical form* of a set of clauses to be the minimal set of clauses that implies *Cn*. The minimal set in the example is $\{\{A\}\}$. There are various ways to compute the canonical form effectively. We will assume that every conjunction is brought into its canonical form before it is retracted.

Conjunctions introduce a complication into the the distance measurement between sets of possible worlds because there is more than one truth-value assignment to the conjuncts that makes a conjunction false. The issue is whether we should count as closest to a set W verifying a conjunction *P all* $\neg P$-worlds that differ from any world in W only with respect to *P*, or whether we should draw finer distinctions among the $\neg P$-worlds, based on the number of conjuncts involved in *P*'s falsification.

For example, let T be the closure of $\{B\}$, $\{C\}$, and $\{D\}$, characterized by W, and let $P = \{\{A, B\}, \{C\}\}$. We are interested in the closest $\neg P$-worlds:

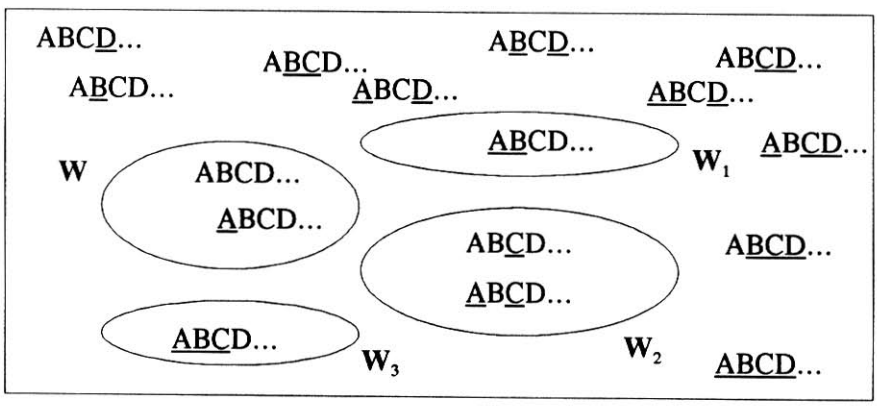

Figure 4

The closeness criterion (C1) is not sensitive to the number of conjuncts involved in the falsification of *P*. It counts as closest any world falsifying *P* that otherwise agrees with some world in W. According to (C1), the set of $\neg P$-worlds closest to W is $W_1 \cup W_2 \cup W_3$ so that $T_{\bar{P}} = W \cup W_1 \cup W_2 \cup W_3$. Intuitively, however, it seems preferable to say that W_3 is farther from W than either W_1 or W_2 because each world in W_3 differs from any

world in W more than any world in W_1 or W_2. This suggests the following more discriminating closeness criterion:

(C2) A non-empty set S^- of $\neg P$-worlds is said to be *closest* to a set of worlds S in which P is valid just in case every member s^- of S^- falsifies exactly one conjunct C_i of P, and there is some member s of S with which s^- agrees in all respects except those falsifying C_i.

According to (C2), the set of $\neg P$-worlds closest to T is $W_1 \cup W_2$, and $T_{\bar{P}} = W \cup W_1 \cup W_2$. Minimality clearly favors this option; I therefore adopt (C2). The earlier proofs remain intact because for clauses in L the two definitions are equivalent.

There is another alternative, or rather class of alternatives: one could set $T_{\bar{P}}$ to only one of the smaller sets $W \cup W_1$ and $W \cup W_2$ and thereby stay even closer to the original theory. It is not clear, however, which of the smaller sets to choose in general because there will be a tradeoff between the size of the sets and the minimal number of atomic sentences by which their worlds differ from the worlds in the reference set. If, for example, the reference set is {ABCD..., A̲BCD...} and we want to retract {{*A*, *B*, *C*}, {*D*}}, then the closest not-{*A*, *B*, *C*} set is {AB̲CD...}, and the closest not-{*D*} set is {ABCD̲..., A̲BCD̲...}. In order to choose, we have to weigh set size against distance of content, and I have no idea how to do this *a priori*. One might decide to go always for the smaller set, which would amount to retracting only the most complex conjunct, because the more complex the conjunct the smaller the closest set of worlds in which it is false. But then one could not, as in the example, withdraw {*A*, *B*} and {*C*} as a conjunction, which can be desirable because retracting {*A*, *B*} and {*C*} separately leaves us with a theory that does not contain {*A*, *B*, *C*}, *i.e.*, a theory that is weaker than the one characterized by the choice of $W \cup W_1 \cup W_2$.

Considerations about the relative merits of the conjuncts have their place in the *preparation* of a contraction, that is in the choice of the object to retract. They should not be taken to constrain the contraction rule itself. By choosing (C2) as the relevant criterion, we do not pretend to have an answer to the Duhem-Quine problem. We do not pretend to know what to do in general when we find one of our theories in conflict with the evidence. We only claim to know what to do —in the simple propositional case— once we have singled out an undesired conjunction and decided not to put the blame on any particular conjunct. (C2) does not assist us in reaching this decision.

(C2) gives rise to the following contraction rule, which treats all conjunct clauses C_i alike:

R2 ($\{C_1, \ldots, C_n\}$, T) If T does not imply all the clauses C_i in P, do nothing. Otherwise apply *R1* recursively to all C_i, *i.e.*, perform (*R1* (C_1...(*R1* (C_n, T))...), and then add the union of every 2-element subset of P to the resulting set.

The unions of the 2-element subsets of P are the strongest clauses we are allowed to add without reintroducing any one of the C_i and thus breaking symmetry; these clauses imply the unions of all larger subsets of P. The trace that remains, for example, of $\{\{A\}, \{B, C\}, \{D\}\}$ is the set $\{\{A, B, C\}, \{B, C, D\}, \{A, D\}\}$. (The rule that would correspond to the less discriminating closeness definition (C1) is *R2* without the second step in which clauses deleted in the first step are added again.) Contraction can be defined as before as the logical closure of *R2* ($\{C_1, \ldots, C_n\}$, T).

R2 preserves what the contractions by the different conjuncts have in common; the effect of adding the union of the C_i is that $T_{\overline{\{C1,\ldots,Cn\}}} = T_{\overline{C1}} \cap \ldots \cap T_{\overline{Cn}}$. In our example, we first contract T by $\{C\}$, which corresponds to joining W and W_2. Then we contract by $\{A, B\}$, which corresponds to joining $W \cup W_2$, W_1, and W_3. Finally, we add $\{A, B, C\}$, which corresponds to subtracting W_3, so that we end up with $W \cup W_1 \cup W_2$, the union of the sets of worlds characterizing $T_{\overline{\{A,B\}}}$ and $T_{\overline{\{C\}}}$. (The order of the contraction steps is irrelevant.)

The two rules *R1* and *R2* form a complete account of revision for logically closed theories in L by sets of clauses in canonical form. Syntactic and semantic approach are equivalent in the following sense:

*Equ** If a set of worlds W characterizes a theory T, and (C2) is the closeness measure for *Con*, then *Con* (P, W) characterizes *Cn* (*R2* (P, T)), where P is a set of clauses in L.

I omit the proof for *Equ**; it is essentially a combinatorial explosion of the proof for *Equ*.

4. *Closure*

It is time now to inject some realism into the discussion and drop the closure requirement on theories. We want to be able to modify the fragments of theories that we actually represent on paper, in computer memories, or in our minds. Contraction and revision should be unaffected by the way the theory is stated; so we postulate that $O \bar{-}_P = (Cn\ O) \bar{-}_P$, for any consistent set of clauses O. If we simply applied *R2* to O, this postulate would be violated in two kinds of case. A clause that is not in O might be irretrievably lost if it is a superset of some clause in O or if it is implied by some subset R of O without being a superset of any of R's elements. The first kind of case is illustrated by $O = \{\{A\}\}$ contracted by $\{A, B\}$. *R1* ($\{A, B\}$, *Cn* (O)) contains $\{A, \neg B\}$, but *R1* ($\{A, B\}$, O) neither contains nor implies $\{A, \neg B\}$. An illustration of the second kind of case is $O = \{\{\neg A, B\}, \{A\}\}$ contracted by A. *R1* ($\{A\}$, *Cn* (O)) contains B, but *R1* ($\{A\}$, O) neither contains nor implies B. Mixed case are of course also possible.

Two different remedies are called for. The remedy for the second kind of case is to bring O into canonical form before contracting or revising it. The first kind is taken care of by the following reformulation of *R1*:

R1' ($\{L_1, \ldots, L_n\}$, T) If T does not imply $C = \{L_1, \ldots, L_n\}$, do nothing. Otherwise replace every clause D that contains a subset C' of C by the set of clauses $D \cup \{\neg L_i\}$, for each of the L_i in $C - C'$.

Notice that if S is a clause containing C, there is no $\{\neg L_i\}$ to conjoin. In this case, S is replaced by the empty set, *i.e.* deleted. We have *R1'* ($\{A, B\}$, $\{\{A\}\}$) = $\{\{A, \neg B\}\}$, as desired. I leave it to the reader to verify that the two remedies work in general.

5. *Entrenchment revisited*

R2 is not without quirks. For example, suppose theory T = *Cn* $\{\{A, B\}\}$ is first expanded by A and T' = T $\dot{+}_A$ then contracted again by A. The result is *Cn* $\varnothing$, although intuitively we may want to restore the initial theory T. Even if we do not think of $\{\{A, B\}, \{A\}\}$ as the result of an expansion by A, we may take the explicit mention of the disjunction $\{A, B\}$ as an indication that there are reasons other than A for its presence, reasons which count

against its removal in the course of removing A. Let me flesh out the example a little. Suppose you believed for good reasons that there is a ball in exactly one of two urns, A or B. Now someone tells you that the ball is in urn A, which you accept. Afterwards you learn that the informant is a notorious liar. You will then reject what you came to infer from his testimony, but if in revising your view you follow *R2*, you will no longer believe that there is a ball either in urn A or in urn B. Yet of course you still want to believe this disjunction. So *R2* appears to give bad advice.

Your reasoning in favor of retaining the disjunction is based on a wealth of tacit assumptions not represented in T'. Most importantly, you think that your belief in the disjunction has independent —and better— grounds than your later acquired belief in one of the disjuncts. This thought can be expressed very naturally in terms of epistemic entrenchment by saying that $\{A, B\} >_{T'} \{A\}$. And the entrenchment ordering can be used in combination with *R2* to yield the desired contraction.

It is easiest to approach contraction through revision. Recall that a total ordering $\leq_T$ over the sentences in L induces a system of nested spheres around T as shown in figure 2. Call the smallest $|\neg A|$-intersecting sphere $S_{\neg A}$ and the set of sentences characterized by it P. The desired revision by $\neg A$ is given, not by all $\neg A$-worlds closest to T, but by those closest $\neg A$-worlds that are also in $S_{\neg A}$. In these worlds, all sentences in P are valid. So the selection amounts to revising T by $P \cup \{\neg A\}$. $T^{*}_{P \cup \{\neg A\}}$ is definitionally equivalent to $(T^{-}_{X})^{+}_{P \cup \{\neg A\}}$, where X is the "negation" of $P \cup \{\neg A\}$, *i.e.*, the disjunction of A and the negations of all the P_i in P, brought into canonical form. T^{-}_{X} can be computed using *R2* in the normal way. (If L were an infinite language, the canonical form of X might not be finite, in which case *R2* $(X, -)$ would not be an effective procedure.)

In the urn example, we wanted to preserve the disjunction $A \vee B$ in the contraction of $T' = Cn\{\{A\}\}$ by A. What we want is $T'^{-}_{\neg(A \vee B) \vee A}$, which is the same as $T'^{-}_{\{A, \neg B\}}$. A quick inspection of *R1'* shows that this theory indeed contains $\{A, B\}$, as desired. In the simplest case, with which we have been concerned in the bulk of this paper, all non-tautological clauses in T were equally entrenched. Thus for any non-tautological A implied by T, the smallest $|\neg A|$-intersecting sphere is the set of all possible worlds. So P is the set of tautologies, and contracting T by A while preserving P is contracting T by $(A \vee \ldots \vee \neg P_i \vee \ldots)$, for all tautologies P_i. This is obviously tantamount to simply contracting T by A.

The virtue of the approach taken in this paper is that it allows us to construct syntactically specified contraction and revision functions from an

arbitrary entrenchment ordering. The ordering can be as simple or as complex as we please, which is how it should be if epistemic entrenchment is a substantive notion, not just a new label for the revision problem. The story is still not complete, though. Entrenchment is theory dependent, and we have not said anything about how to revise *it* in the course of theory revision. We therefore cannot yet handle iterated theory revision.

CREA / Ecole Polytechnique
LIPN / Université Paris-Nord

REFERENCES

Alchourrón, C. E., P. Gärdenfors, and D. Makinson (1985). "On the Logic of Theory Change: Partial Meet Functions for Contraction and Revision." *Journal of Symbolic Logic* 50: 510-530.

Gärdenfors, Peter (1988). *Knowledge in Flux. Modeling the Dynamics of Epistemic States* (Cambridge, Mass. & London: MIT Press).

Gärdenfors, Peter, and David Makinson (1988). "Revisions of Knowledge Systems Using Epistemic Entrenchment," in *Theoretical Aspects of Reasoning about Knowledge*. Ed. Moshe Y. Vardi (Los Altos, CA: Morgan Kaufmann): 83-95.

Genesereth, Michael R., and Nils J. Nilsson (1987). *Logical Foundations of Artificial Intelligence* (Los Altos: Morgan Kaufmann).

Grove, Adam (1988). "Two Modellings for Theory Change." *Journal of Philosophical Logic* 17: 157-70.

Harman, Gilbert (1986). *Change in View: Principles of Reasoning* (Cambridge, Mass.: MIT Press).

Levi, Isaac (1984). *Decisions and Revisions. Philosophical Essays on Knowledge and Value* (Cambridge etc.: Cambridge University Press).

Levi, Isaac (1989). "Review: Knowledge in Flux. Modeling the Dynamics of Epistemic States." *The Journal of Philosophy* LXXXVII. 8: 437-444.

Lévy, François (1993). "Revision in Classical Logic." *Unpublished paper.*

Lewis, David K. (1973). *Counterfactuals* (Cambridge, Mass.: Harvard University Press).

Mongin, Philippe (1993). "The Logic of Belief Change and Nonadditive Probability," in *Proceedings of the 1991 LMPS Conference*. Ed. Dag Prawitz and Dag Westerståhl (Dordrecht: Kluwer, Synthese Library).

Papini, Odile (1992). "A Complete Revision Function in Propositional Calculus," in *ECAI 92. 10th European Conference on Artificial Intelli-*

gence. Ed. B. Neumann (Chichester etc.: John Wiley & Sons).
Rott, H. (1991). "Two Methods of Constructing Contractions and Revisions of Knowledge Systems." *Journal of Philosophical Logic* 20: 149-173.