Changing the Theory of Theory Change: Towards a Computational Approach

Neil Tennant

The British Journal for the Philosophy of Science, Volume 45, Issue 3 (Sep., 1994), 865-897.

Stable URL:
http://links.jstor.org/sici?sici=0007-0882%28199409%2945%3A3%3C865%3ACTTOTC%3E2.0.CO%3B2-Q

Brit. J. Phil. Sci. **45** (1994), 865–897

Changing the Theory of Theory Change: Towards a Computational Approach

Neil Tennant

ABSTRACT

The theory of theory change has contraction and revision as its central notions. Of these, contraction is the more fundamental. The best-known theory, due to Alchourrón, Gärdenfors, and Makinson, is based on a few central postulates. The most fundamental of these is the principle of recovery: if one contracts a theory with respect to a sentence, and then adds that sentence back again, one recovers the whole theory.

Recovery is demonstrably false. This paper shows why, and investigates how one can nevertheless characterize contraction in a theoretically fruitful way. The theory proposed lends itself to implementation, which in turn could yield new theoretical insights. The main proposal is a 'staining algorithm' which identifies which sentences to reject when contracting a theory. The algorithm requires one to be clear about the structure of reasons one has for including sentences within one's theory.

1 The problems of theory contraction and revision

> Given a theory T, and a statement S therein, how can one simply give up commitment to S while holding on to as much of T as possible?

This is the problem of *contraction*, and it will be the sole concern of this paper. Surprisingly, it is only relatively recently that logicians have addressed the problem in a general and rigorous way.[1] That the problem is indeed a central one can be appreciated by noting that its solution would provide also a solution to the problem of how to *revise* a theory T. If T commits us to S, and we wish instead to believe $\sim$S, the way to get the desired result T*$\sim$S (the revision of T with respect to $\sim$S) would be first to contract T with respect to S and thereafter to expand the contracted theory (T-S) by adding $\sim$S:

$$T^*\sim S = [T\text{-}S, \sim S], \text{ where } [\ldots] \text{ indicates logical closure.}$$

This is known as the Levi identity.[2] If we have done our job correctly in contracting with respect to S, then the subsequent expansion with respect to $\sim$S will yield a consistent result.

Like other writers on these topics I endorse the Levi identity. But I do not endorse all that has been written about contraction. Alchourrón, Gärdenfors, and Makinson have together developed a significant body of theory[3] on the logic of theory change. They proceed as follows. They lay down some basic and global postulates involving contraction, expansion, logical consequence, logical closure, and set-inclusion. They also invoke an *entrenchment relation* (more on which in Section **2.2** below) which registers, roughly, one's strength of commitment to the various statements in one's theory. These postulates provide constraints on the construction of explicit contraction functions T-S, for any theory T (with its pattern of entrenchment) and any sentence S. Their aim is to establish representation theorems of the form 'Contraction satisfies such-and-such global postulates if and only if it is given by a function defined thus-and-so by appeal to

[1] The AI community made a start with their so-called 'truth-maintenance systems' (TMS) or 'reason-maintenance systems' (RMS). See Doyle [1979] and de Kleer [1986]. But the virtues of these earlier approaches have perhaps been neglected because a fully developed metalogical theory was initially lacking.

[2] The Levi identity has its source at p. 427 of 'Subjunctives, Dispositions and Chances' [1977].

[3] See the references below. This theory is now commanding serious attention within the AI community. See, for example, the collection edited by Fuhrmann and Morreau [1991]. A cause for possible regret is that in buying this defective theory from the logicians, the AI community may stand to lose their earlier crucial insights (within TMS and AMS) about the structure of justifications within theories. What we need to do is develop those earlier insights into a better metalogical theory, and, one hopes, not sacrifice implementability in doing so.

the entrenchment relation [among (sets of) sentences] with such-and-such features'.[4]

My contention is that there is not very great value in such high-level theoretical results if the contractions so constrained and so defined deliver non-viable results in simple cases. I would therefore rather spend more effort getting clearer about the nature of contraction, before seeking formal representation theorems of this kind. This essay takes the form, then, of an extended misgiving about the foundations of an already impressive edifice in whose shadow I am constrained to make my way. Let me now foreshadow what I shall be putting in its place, and explain briefly what is novel about the approach I shall develop. In doing so some criteria of adequacy will emerge.

I should stress at the outset that it is not my intention here to burden the reader with unnecessary technical definitions; nor do I intend to furnish technical results. The technical results will be published separately. The intention in this paper is to reach as wide an audience as possible—philosophers of science, logicians, and workers in AI—by keeping technicalities to a minimum. Sometimes the main shortcomings of an orthodox position, and the leading ideas of an opposing heterodox position, can be spelt out very cleanly with a minimum of symbolism. This is the case with the orthodox theory of theory change, and with the radically changed theory of theory change into which I would have it changed. The inadequacies of the orthodox theory, and the unsatisfactory state of the secondary literature on it, invite a harder look at the problems.

2 From old to new

2.1 Recovery

The main foundation stone of the Alchourrón–Gärdenfors–Makinson theory (AGM) is what Gärdenfors calls the principle of *recovery*:

> If one contracts a theory T with respect to one of its consequences S, and then assumes S alongside the result, one can deduce all of T. That is, what one gets rid of by contraction with respect to S should not exceed S in deductive power within T.
> Formally: $T \vdash S$ and $T \vdash R \Rightarrow T\text{-}S, S \vdash R$

I shall be arguing that this principle has to be abandoned in the light of

[4] This holds true even of Hansson [1989], who has on p. 123 his main theorem stating that a contraction operator satisfies his 'extended' Gärdenfors postulates iff it is what Hansson calls an operator of composite partial meet contractions. The methods adopted by Hansson in response to the perceived limitations of the original approach are not substantially different from those of the Alchourrón–Gärdenfors–Makinson theory.

intuitively forceful counterexamples that leave us no alternative but to ditch it. The counterexamples are given in Section **3** below.

2.2 Entrenchment

Entrenchment[5] can be thought of as a pre-ordering $\leqslant$ of sentences in the theory. $S \leqslant R$ means 'R is at least as well entrenched as S'. A pre-ordering is a relation that is reflexive and transitive. Our pre-ordering $\leqslant$ moreover respects deducibility, in the sense that if R is deducible from S then $S \leqslant R$. A consequence of this requirement is that for logically equivalent sentences S and R we have both $S \leqslant R$ and $R \leqslant S$. The strict relation $S < R$, 'S is less entrenched than R', can then be defined as $(S \leqslant R \ \& \sim R \leqslant S)$. This relation is asymmetric, transitive, and non-circular. The relation $<$ need not be connected.[6] Thus its field could decompose into separate 'islands' of sentences that are all linked into a mesh of entrenchment comparisons, while yet the islands are incomparable. Indeed, there could be isolated elements in the ordering $<$, sentences that cannot be said to be more or less entrenched than any others. By the usual convention, such elements would count as $<$-minimal. Thus if, say, a three-element premiss set $\{A,B,C\}$ were culled from a theory, within which the entrenchment facts were simply that $B < C$ (and hence that A was isolated), then the $<$-minimal members of this set would be A and B. Thus 'x is $<$-minimal' means 'for no y is it the case that $y < x$'.

2.3 Implementability

The main results of the AGM theory are representation theorems to the effect that the operation of contraction satisfies certain global postulates if and only if it can be defined in a particular way, using the resources of set theory and metalogic, and the available facts about entrenchment. Two quick observations are in order:

(1) The method of definition does not lend itself to implementation in a way that would be useful in AI.
(2) Among those postulates is recovery.

In later work,[7] Gärdenfors has tried to meet objection (1). He appeals to the facts about entrenchment[8] to enable one to 'compute' the contraction of a theory T with respect to (contingent) S by defining the contraction as

[5] For these purposes, entrenchment was introduced by Gärdenfors [1984]. The converse of the relation of (strict) entrenchment appears in Fuhrmann [1991] under the name 'retractability'. Fuhrmann does not seem to be aware of this.
[6] Here we differ from Gärdenfors [1984].
[7] See Gärdenfors [1990].
[8] More on this below.

consisting of exactly those members R of T such that the disjunction R ∨ S is strictly more entrenched than S. Such contraction satisfies all the usual postulates, including recovery.[9] And *that*, I say, is why Gärdenfors' recent nod in the direction of the demands for implementability still will not do. For recovery has to go. And the examples below that show why it has to go cannot be massaged into conformity with Gärdenfors' account by rigging the entrenchment relation before applying his methods.

For that reason also (the failure of recovery), Fuhrmann's [1991] treatment of 'minimal' contractions of finite bases has to go.[10] For, as Fuhrmann himself concedes (p. 179), when the base is a theory, his account of 'minimal' contraction coincides with Alchourrón and Makinson's account of so-called 'safe' contraction; and the latter validates the principle of recovery for theories.[11]

So we are still looking for an implementable theory that (a) is not committed to the principle of recovery, (b) is implementable, or at least a source of fruitful computational models of theory-change, and (c) applies quite generally to sets of sentences, whether or not they are logically closed. No author has so far met all these desiderata. In this paper I want to suggest how we might do so.

The treatment in Gärdenfors [1990], in so far as it shows an interest in computational concerns, tries to get all its computational purchase via the epistemic entrenchment relation. In particular, Gärdenfors tries to recover the *structure of internal justifications*, or *reasons* for our beliefs, solely from facts about their relative entrenchment. I shall show that this is unduly restrictive. It is arguably descriptively inadequate, and most certainly *normatively* inadequate.

A simple example will make clear the need to keep separate accounts of the structure of *justifications*, on the one hand, and the structure of *entrenchment*, on the other. Consider a well-founded belief system with various basic assertions forming its evidential basis, and with a relation of support between sets of premisses and conclusions that is not necessarily strictly deductive, but is nevertheless still monotonic.[12] That is, we are equipped with the notion, for the language of such a system, of (basic)

[9] See Gärdenfors and Makinson [1988].

[10] For an independent criticism of the shortcomings of Fuhrmann's allegedly 'minimal' contractions, see below. These contractions can, in general, *maximize* the mutilation inflicted on a theory that is to be contracted.

[11] See Alchourrón and Makinson [1985]. Fuhrmann [1991] in effect perseveres with their mistake of getting rid of *all* so-called 'unsafe' sentences, instead of picking unsafe sentences off one at a time. But in their defence, Alchourrón and Makinson, unlike Fuhrmann, were not claiming to be minimal mutilators in doing so.

[12] The assumption of monotonicity made just for the purposes of this example is not required in general within my treatment. In general, I allow for non-monotonic relations of justification between sets of premisses and conclusions.

premisses $P_1, \ldots, P_n$ supporting a conclusion Q to which they are relevant, and continuing to do so even when extended by further basic sentences $P_{n+1}, \ldots, P_m$. When such a conclusion (belief) Q has to be given up, the question arises as to which of the P_i we should also give up. Now the various P_i could very well be logically and evidentially independent of each other. Thus there is no structure of *reasons for believing* to be discerned among the P_i themselves. Yet there could still be a structure of relative entrenchment among the P_i that would make it possible to choose some of them rather than others for rejection along with Q. This could well be because of the believer's preference (albeit not registered or expressed explicitly *within the system*) for evidence from one modality (such as sight) rather than another (such as smell). In such a case, *both* the structure of justifications (that is, the inferences drawn among the beliefs, which lead to the adoption of certain Qs rather than others, 'on the basis of the evidence') *and* the pattern of relative entrenchment importantly come into play in furnishing an answer to the question of which beliefs to surrender in the wake of an ousted Q. This is but one small example, but the general point it makes is incontrovertible. We should neither conflate the pattern of entrenchment with the pattern of justification, nor seek to reduce, as Gärdenfors does, the one to the other.

Gärdenfors is not clearly enough distinguishing what is in his model and what might be in the phenomena of which it is supposed to be a model. By contrast, the approach I favour will exploit the full structure arguably available to the theorist who is sensitive both to the justificatory pedigrees an agent may have for her beliefs, and to facts about those beliefs' relative entrenchment within her system of beliefs, *without* committing itself to a reduction of either one of these structures to the other. So I shall have two dimensions within which to operate, instead of only one.

3 A new perspective

The attention that I am recommending we should pay to the structure of justification within a theory enables us to attain a new perspective on the problem of theory contraction. The problem has so far been considered by other theorists only from what I shall call the *limited perspective*. This is the point of view according to which the sole purpose of contraction is to get rid of the sentence S *as something that might be justified* within the theory T. But the limited perspective ignores the need for whatever corrections may be called for when we realize that, within the theory T, that sentence S (and others to be expelled with it) might have furnished indispensable *justifications for other sentences* of the theory

T.[13] We might say that from the limited perspective one looks only *upwards* along lines of logical dependence. One looks upwards from S as a *conclusion* of proofs or justifications to the premisses thereof, those premisses being also in T. And one tackles the question of which of these premisses one is to throw out along with S when contracting T with respect to S. The relation of (strict) entrenchment is invoked to help with this choice of vulnerable or culpable S-justifiers: from the set of premisses of any perfect proof of S (that is, a proof of S whose axiom set has no proper subset implying S) one strikes out one least entrenched (that is $<$-minimal) member. (The best way to do this will be considered later.)

I want now to recommend another, complementary perspective, which seems to have been neglected by other writers on this problem. This new perspective might be thought of as involving looking *downwards* along lines of logical dependence. One should look down from S as a *justifying premiss* to conclusions R in T that *rest essentially* on S or on other sentences to be dropped, in the sense that every proof in T of R draws on S or on one of those other sentences as a premiss.[14]

On the limited perspective one looks upwards at progenitors of S. On the new, complementary perspective I am recommending, one also looks downwards at the progeny of S and of other exiles. On the limited perspective we see S and its progenitors (enough of these, chosen via entrenchment) as liable for exclusion. On the new perspective *we see the totally dependent progeny of S and other exiles as liable for exclusion also.*

An immediate casualty, then, is the Gärdenfors postulate of recovery:

> If one contracts T with respect to one of its consequences S, and then assumes S alongside the result, one can deduce all of T. That is, what one gets rid of by contraction with respect to S should not exceed S in

[13] Hansson [1989] points out on p. 118 that it is counterintuitive for the contraction of [A] with respect to A to imply any disjunction $A \vee B$, but that if the original theory provided independent reason for believing the disjunction, then $A \vee B$ *should* survive the contraction with respect to A. Hansson's own proposals, however, remain squarely within the conceptual and technical resources of the original AGM theory. He reverts to a treatment in terms of finite bases, which is an over-reaction to the problem. What is needed, instead, is the more sensitive approach to the representation of intra-theoretical dependencies that I shall be developing below. A similar criticism applies to Fuhrmann [1991].

[14] There is an undeveloped and to my mind confused hint of this idea in Martins [1991], at p.60. But so little is said that it is not clear whether by 'disbelief propagation' Martins means 'propagation of suspension of belief'—*i.e.* contraction. If so, it is terminologically inapt, and his paper does not show one how to go about it. If not, then the following statement of his is a howler: 'The result of this disbelieving process should be the set of propositions that would have been generated if the system had started without the disbelieved proposition.' To see that this cannot be right, take the single contingent statement A, and consider its logical closure [A]. Now change your mind, and disbelieve A. That means your new theory is $[\sim A]$. But the system that would have been generated had you started without the disbelieved proposition A is [Ø], that is, pure logic alone.

deductive power within T.
Formally: $T \vdash S$ and $T \vdash R \Rightarrow T\text{-}S, S \vdash R$

The reason why recovery fails, given the new perspective, is as follows. Consider the theory $[A, A \supset B]$ where $A < (A \supset B)$. Its contraction with respect to B will contain $A \supset B$ but not A. Before the contraction the theory $[A, A \supset B]$ would also have contained $B \supset A$, which on the limited perspective would have survived the contraction, thus serving, upon expansion with respect to B, to precipitate A. But on the new perspective the sentence $B \supset A$ might be regarded as totally dependent progeny of A. In such a case the contraction would go:

get rid of B;
thus choose to get rid of one of $A, A \supset B$;
since $A < (A \supset B)$, it will be A that has to go;
now get rid of the totally dependent progeny of A;[15]
hence get rid of $B \supset A$...

Now, when we expand with respect to B, the contracted theory will not have $B \supset A$ present to help precipitate A back again. Summarizing the foregoing, we have:

$$[A, A \supset B] \vdash B \text{ and } [A, A \supset B] \vdash A; \text{ but not } [A, A \supset B]\text{-}B, B \vdash A,$$

which is a clear counterexample to the recovery postulate

$$T \vdash S \text{ and } T \vdash R \Rightarrow T\text{-}S, S \vdash R.$$

What is notable here is that we are still dealing with *theories*, albeit theories with some extra structural information about justificatory pedigrees.[16] To wit, we had the information that the justification for $B \supset A$ in the original theory was that the theory contained A.

The normative demands of the new perspective that I am recommending are not easy to accommodate when theories are regarded simply as (logically closed) sets of sentences, without any of this extra structural information about justificatory pedigrees. It could turn out, for example, that some sentence D is in the theory T only because F is; and that F is one of those worst entrenched progenitors of E targeted for exclusion in order to avoid having E in the result. Take, for example, the theory [A,B], where A is more entrenched than B. This theory contains neither C nor $\sim C$, but it

[15] This is where the new perspective makes a difference that will be felt.
[16] Note that in Makinson [1987] it is conceded that recovery fails when contraction and expansion are performed on a finite (hence not logically closed) set of sentences; but that recovery holds for theories. What I have shown is that recovery should fail to hold in general even for theories, provided only that they are conceived as involving not only logical closure but also information about justificatory pedigrees of each theorem. The conceptual consequences will be amplified below.

does contain B ∨ C. If we wish to contract the theory with respect to A&B we choose to get rid of B because B is less entrenched than A. Heeding the maxim of minimum mutilation, however, other writers would leave B ∨ C in the contracted theory, since it plays no part in implying the unwanted A&B. But this jars with the demands of intuition. Intuitively, shouldn't we also get rid of B ∨ C? For B ∨ C does, after all, owe its presence in the original theory entirely to the presence of B, which is now about to be banished. Something is not quite right here with the process of contraction as construed from the limited perspective.

Someone might object that whoever held the theory above might, for peculiar reasons, wish to retain commitment to B ∨ C even if they ceased to believe B. In response to this objection we could make the problem even more focused and compelling, and at the same time more degenerate. (By 'degenerate' one means here 'free of any distracting ingredients'.) Take the theory [B,∼C] where B is more entrenched than ∼C. We note that B ∨ C is in this theory. We note also that B ∨ C, ∼C ⊢ B. Imagine now the contraction of this theory with respect to B. Clearly one of B ∨ C, ∼C must go—preferably the less entrenched of these, if they are not equally entrenched. But which is the less entrenched? Intuitively, any disjunction B ∨ C must be at least as well entrenched as its disjunct B. And we are here assuming that B is more entrenched than ∼C. By transitivity of entrenchment, B ∨ C is more entrenched than ∼C. So it is ∼C that must go, leaving B ∨ C behind in the contracted theory. The result of the contraction is therefore [B ∨ C].

But this is wrong. Nothing could be clearer than that in the theory [B,∼C] the sentence B ∨ C is present *only because B is*. (B ∨ C couldn't be following from C, because, on pain of contradiction, C is absent from the theory.) Hence, when we contract with respect to B, the sentence B ∨ C should go *immediately*, without so much as pausing to have a look at the role of progenitors of B within the theory. Now, if we heed this intuition and get rid of B ∨ C, the result of the contraction is [∼C]. This is surely the correct result.

The principle of minimum mutilation, for theories conceived of as (logically closed) sets of sentences, militates against our following such intuitions through. It enjoins us to retain as much as we can of the disjunctive progeny of sentences being excised, provided only that those disjunctions play no vulnerable role in implying any sentence being excised. (Vulnerability is determined by entrenchment in a way that is now clear.) Thus, in accordance with minimum mutilation, and with B more entrenched than ∼C, the limited perspective would have it that [B,∼C] − B (the contraction of [B,∼C] with respect to B) would imply (hence, as a theory, contain) B ∨ C. But, on the new perspective we are considering, it should not.

Though we have here had to curb the excesses of unbridled commitment to minimum mutilation, in the interest of avoiding clearly counterintuitive results when we attend to the structure of justifications within a theory, the reader should not conclude from that fact that we are ready to jettison the requirements of minimum mutilation altogether. We certainly are not. It remains as one desideratum among others; combined with them, it provides a creative tension within which the contraction of a theory will be fashioned. We might modify the statement of the principle of minimum mutilation so that it reads:

> Theories should be contracted in such a way as to effect minimum mutilation, *compatibly with the result not having unjustifiable excrescences.*

We have been considering examples where a sentence (like $B \vee C$) is not a culpable progenitor of anything that has to go, but is rather deprived of any right to remain on the contracted scene once its own progenitors have been expelled. Indeed, my own intuition is that this consideration should be able to *pre-empt* ones to the effect that the sentence concerned might be to blame for the presence of some other unwanted sentence(s).

I have raised my examples as a rather serious problem for the account of contraction that we have from the work of Gärdenfors, Makinson, and Alchourrón, and, so far, of all their commentators. Looking only at culpable progenitors and relative entrenchment can produce the wrong answer. To get the right answer we need to look also at progeny that will inherit a stain, and get rid of them early.

4 Foundationalism vs Coherentism

In his anxiety to commend the supposedly 'computation-friendly' version of his theory,[17] Gärdenfors seeks to ally his approach with that of the coherentist tradition in epistemology. He thinks that interest in justificatory pedigrees is the exclusive concern of the foundationalist, and that interest in minimal mutilation of theories, or in minimum loss of information (on contraction or revision of our theories), is the exclusive concern of the coherentist.

Nothing could be further from the truth. Both the foundationalist and the coherentist are (or should be) concerned both with justificatory pedigrees and with minimal mutilation and minimal loss of information. I claim it as a virtue of my approach below that it allows one to describe (and prescribe) a method of contraction (hence of revision) that would be available equally to the foundationalist and the coherentist. The difference

[17] Gärdenfors [1990] is the paper in question here.

between them would lie solely in the respective 'topologies' of the system of justificatory pedigrees. (For the foundationalist, but not for the coherentist, there will be a requirement of well-foundedness.) But the *method* of contraction is invariant across the disagreements between the foundationalist and the coherentist, and is indifferent to the outcome of the epistemological debate between them. The new perspective is not a mere retreat to the idea, put forward by some writers,[18] that we should perform contractions on closures of *bases* for theories. For representing theories as closures of bases is to commit oneself to a foundationalist perspective, according to which, presumably, sentences in the base will be more entrenched than their distant consequences (if considerations of entrenchment are still in play). Yet the new perspective applies not only to theories conceived in a foundationalist fashion, but also to theories conceived in coherentist fashion. Indeed, one of the consequential virtues of this point is that the application of my method (since it is normative) could help one generate materials on theory-contraction that could help adjudicate the debate between the foundationalist and the coherentist. This is because the algorithmic method, once implemented, would provide a powerful tool for spelling out the ineluctable consequences, in the case of various example theories or types of theory, of being true to one's foundationalist or coherentist convictions and at the same time undertaking contractions and revisions in the manner in which, I contend, one indifferently ought.

5 On the internal structure of theories

To this end it is important to have built into the constitution of our theories an account of how, by the theory's own lights, its sentences come to be there.

Might one not object, however, that this information on pedigrees of presence is already there, available and waiting to be 'read off' from the mere facts of sentence membership in the theory? For example, a theory that contains B and $\sim$C can contain B $\vee$ C *only* by virtue of B's presence in the theory. Isn't this just a small example of how every sentence in a theory might be uniquely (or multiply) pedigreed?

I think the answer to this rhetorical question is 'No'. As conceded in the objection raised above, a theory $\{\ldots B, B \vee C, \ldots\}$ *not* containing $\sim$C might very well have B $\vee$ C there 'in its own right', so to speak, and *not* as derivative from B. With such a theory the disjunction B $\vee$ C might even survive contraction with respect to B! Although B $\vee$ C does follow logically from B, the point is that it need not be seen as thus depending on B *for*

[18] See, *e.g.*, Hansson [1989], Gärdenfors [1990], and Fuhrmann [1991].

its presence within the theory. This is a situation to be distinguished from one where $B \vee C$ does simply depend on B for its presence in the theory. Yet such a theory (even one not containing $\sim C$) could also be rendered as $\{\ldots B, B \vee C, \ldots\}$ if all we pay attention to is sentence membership.

What we therefore need is some richer articulation of our conception of a theory. This articulation would have to show, for each sentence S in the theory, why it was that S was indeed in the theory. How might this be done?

One way would be to treat as members of the theory not just single sentences S but rather items of the form $(S, \Pi_1, \ldots, \Pi_n)$ where the $\Pi_1, \ldots, \Pi_n$ are all the severally sufficient intra-theoretic 'pedigrees' of S.[19] We could think of them as *proofs*, *within the theory*, of S. Thus one might have an entry like

$$\left(B \vee C, \frac{B}{B \vee C}, \frac{C}{B \vee C}\right)$$

for the disjunction $B \vee C$ in the theory that would normally be described as [B,C] on the understanding that B and C were its axioms, and not themselves requiring any justification. (An axiom A might be thought of as taking the form (A,[A]).)

Instead of $(S,\Pi_1, \ldots, \Pi_n)$ one might use $(S,\Delta_1, \ldots, \Delta_n)$, where each Δ_i is the set of premisses of Π_i. Indeed, given the possibility of multiple proofs of S from one and the same set of premisses, some of the premiss sets in the sequence $\Delta_1, \ldots, \Delta_n$ might be identical. In this case, repetitions could be deleted without loss.

It may be that most entries would be of the form (S,Π)—involving a *single* justificatory pedigree Π. This would be fine; it would make all the easier our task of identifying, as the need arose, those progeny S inheriting a stain from the axioms used in Π. For S with multiple pedigrees $\Pi_1, \ldots, \Pi_n$, S would inherit a 'downward' stain if and only if each of the respective axioms sets $\Delta_1, \ldots, \Delta_n$ had a stained member. (They need not have a stained member in common; it is enough just to disable each Π_i by staining one of its premisses.)

One could impose further global conditions of coherence and foundation on the assignment of pedigrees to sentences, according to one's epistemological tastes. For the foundationalist it would not do, for

[19] Cf. Doyle [1979], de Kleer [1986], and Martins [1991]. This proposal is not as drastically crude as the proposal that we should identify theories simply via their *bases*: finite sets of sentences that are not logically closed but whose closures are the theories in question. Note that my proposal is not exactly like Doyle's. On my account there is no need for, and no use made of, anything like Doyle's 'outlist' of (sentences at) nodes. Note also that my proposal is more general than that of Martins. Martins allows, in my terminology and notation, only for one pedigree Π for any claim within a theory. He speaks throughout of 'the origin set' of any claim, as though any claim will have only one way of being justified or proved within a theory. But this is wrong and represents an unnecessary constriction.

example, to have the two sentences A and B in a theory represented respectively by the items

$$\left(\mathrm{A}, \frac{\mathrm{B} \quad \mathrm{A} \equiv \mathrm{B}}{\mathrm{A}}\right) \quad \text{and} \quad \left(\mathrm{B}, \frac{\mathrm{A} \quad \mathrm{A} \equiv \mathrm{B}}{\mathrm{B}}\right).$$

Clearly this would be circular: each of A and B is here represented as buttressing the other via their equivalence! The foundationalist would need to lay down a condition along the following lines:

> Let us define the relation 'H precedes G in the order of justification' as the ancestral of the relation 'F immediately precedes G in the order of justification', where the latter relation holds whenever the theory contains an entry of the form
>
> $$(G, [\ldots \frac{\Delta}{\mathrm{G}} \ldots]) \text{ where F is in } \Delta$$
>
> Then we can require that justificatory precedence be well founded: that is, there should be no loops or infinitely descending chains in the order of justificatory precedence.

This would commit us to a foundationalist picture. But I want, in what I say subsequently, to be able to accommodate a coherentist alternative to the foundationalist picture. I shall therefore try to avoid making use of the foundationalist assumption in what follows. In particular, when I come to frame an algorithm for staining the sentences we have to get rid of, I shall want to ensure that it will terminate, and produce the right result, in the coherentist context just as much as in the foundationalist context.

There would already have been an advantage in conceiving of theory entries in my suggested fashion even on the earlier, limited perspective that concerned itself solely with removing progenitors of a stained sentence without worrying about ensuring the elimination of stained progeny. If the proofs $\Pi_1, \ldots, \Pi_n$ in an entry $(S, \Pi_1, \ldots, \Pi_n)$ are *perfect* proofs (ones whose axiom sets have no proper subsets implying S), then we can target the least entrenched members of the respective axiom sets $\Delta_1, \ldots, \Delta_n$ for removal if S has to go.

With our newly suggested perspective, the Π_i can play a dual role. First, as just noted, they can be used to weed out the most vulnerable of those progenitors that can be blamed for the presence of S. But now, secondly, they can be used to see which sentences would *inherit* the stain of the weedkiller once it has been applied.

6 The staining algorithm

Let us call a proof stained if and only if it has a stained premiss; and let us

call a sequence or set of proofs stained if and only if each of them has a stained premiss (not necessarily the same one in each proof). The process of contraction can then be thought of as follows. The sentence with respect to which we are contracting is stained. The stain is then spread in two ways (to be explained presently) until it can spread no further. Then all stained sentences are removed. The result is the desired contraction. The two ways of spreading stain are called Upwards and Downwards.

> ALGORITHM FOR STAINING:
> Upwards: Find all $(S,\langle\Pi\rangle)$ such that S is stained and at least one member of $\langle\Pi\rangle$ is unstained. Stain exactly one least entrenched premiss of each *unstained* member of $\langle\Pi\rangle$ in a globally sensible way (the sense of which will be explained presently).
> Downwards: Find all $(S,\langle\Pi\rangle)$ such that S is unstained and $\langle\Pi\rangle$ is stained. Stain all such S.

In Upward staining the emphasis on *unstained* members of $\langle\Pi\rangle$ is to prevent one from staining too much. Any stained proof will be disabled anyway when its stained premisses are thrown out in the final throes of contraction. Thus its conclusion will certainly not be able to get back in via that proof. So we do not have to worry about staining any more of its premisses, even if one of its unstained premisses happens to be less entrenched than any of the stained premisses. So what we need to attend to are the *unstained* proofs of S, and make sure that in each such proof we stain one least entrenched (<-minimal) premiss, so that the proof will be disabled.

Now what is meant by a 'globally sensible way' in the step called Upwards? The aim is to disable or cripple each justification in the list $\langle\Pi\rangle = \Pi_1, \ldots, \Pi_n$ while inflicting the least mutilation possible. Each justification Π_i is disabled by having at least one member of its premiss set Δ_i stained. It would be nice if we could achieve all of the following:

(i) in each Δ_i *exactly one* member is stained, and it is moreover a least entrenched member of Δ_i;
(ii) across the $\Delta_1, \ldots, \Delta_n$ as few distinct sentences as possible are stained;
(iii) the stained sentences across $\Delta_1, \ldots, \Delta_n$ are, overall, sprinkled as 'low down' as possible in the entrenchment pattern on $\cup_i \Delta_i$.

Why (i), (ii), and (iii)? Because, in the context of an entrenchment relation, they give expression to the requirement of minimum mutilation. That requirement is, of course, still tugging in the background at all we propose to do. We have to accord it as much weight as possible, compatibly with not committing the gross errors of the limited perspective (those of persevering with completely unsupported beliefs). If ever the principle of minimum mutilation appears to have been compromised, this will only be

because it stands *in tension* with the requirement that we should avoid grossly counterintuitive consequences when contracting a theory. Having made this much clear, we now see why in (i) we call for *exactly one* least entrenched member of Δ_i to be stained, rather than *all* of them, should there be more than one. Staining all of them in the latter case would be to mutilate too much. The structure available in perfect proofs and the preferences registered in the entrenchment relation would be riches down the drain if we resorted to that crude a measure.[20]

7 The *en bloc* approach vs the 'single sentence at a time' approach

I said above that staining *all* of the least entrenched members within the premiss set of a perfect proof of a stained conclusion would be to mutilate too much.[21]

There will in general be a variety of answers to the question 'What is the result of contracting the theory T with respect to the sentence S?' even when T is given complete with justificatory pedigrees. The definite description 'the result . . .' is misleading. There is a unique result only if one insists on a determinate method for making definite choices at each stage when

[20] The mistake of thinking that anyone who appeals to epistemic entrenchment in order to get help with the process of contraction will have resort to it constantly appears to be at work in an otherwise unaccountable claim by Nebel [1990], at p. 162. After noting that if y is a logical consequence of x then y is at least as epistemically entrenched as x, Nebel goes on to say:

> If we have to choose [between] a proposition x and its consequence y, we had better give up x alone instead of y and its *generating* proposition x! In a nutshell, epistemic entrenchment runs counter to the idea of reason maintenance.

This is a serious error. We only ever appeal to epistemic entrenchment to choose culprits (for staining) within the premiss set of a *perfect* proof. *A fortiori*, we shall never have to look at the relative entrenchment of such x and y as Nebel had in mind. For, if x logically implies y, then they cannot both occur as premisses of a perfect proof. This is because we could drop y from any premiss set containing x and still have enough logical power in the remaining premisses of the original proof (among them x) to secure its conclusion. Hence the proof in question could not have been a perfect proof. What this essay sets out to show, among other things, is that, *pace* Nebel, epistemic entrenchment can be recruited both in the service of reason maintenance and in the interests of minimal mutilation!

[21] The mistake against which I am cautioning here is so obvious that such explicit caution might strike some readers as unnecessary. Not so, however: see Fuhrmann [1991]. Fuhrmann defines what he calls a *minimal* base contraction via a reject set (p. 179). The technical definitions of reject set and of minimal contraction embody a blunder, which is masked in the transition to the technicalities. By the expression 'Δ is an *entailment set* for S', Fuhrmann means that the sequent $\Delta : S$ is perfectly valid: that is, it is logically valid, and its premiss set Δ contains no proper subset Γ such that the sequent $\Gamma : S$ is logically valid. (The notion of perfect validity was defined in Tennant [1984].) On p. 178 Fuhrmann had written, concerning contraction of a theory (or its base) with respect to a sentence S:

> We shall have retracted enough sentences from [the finite base], if we subtract *at least one* element from each entailment set for [S] (in [the finite base]). And we shall have done the *least* damage to [the finite base], if we subtract from each

more than one choice is possible. So, if it's just a practical result one is interested in, one could implement that extra dimension of arbitrariness in an automated theory contractor, let it run, and take home the result. But, if one has a theoretical interest in the sheer variety of competing answers that one might get to the question 'What is the result of contracting the theory T with respect to the sentence S?', then one can exploit the power of the computer, with a good implementation of the staining algorithm, to generate a whole spectrum of possibly different outcomes, depending on *which* arbitrarily specific choices for staining were made at each pass of the Upwards step of the algorithm.

If, on the other hand, one does not allow oneself to make specific single choices from among several $<$-minimal candidates at a time, and instead insists[22] that *all* $<$-minimal candidates be stained *en bloc* at every Upwards pass,[23] this could, depending on the structure of the entrenchment relation, produce drastically over-pruned contractions. It would deprive us of the chance that we would otherwise have of producing interesting classifications of contracted theories, and gaining further insights into how best

---

> S-entailment set (in [the finite bases]) *its most expendable* members, i.e. those sentences that are minimal under the ordering $<$. (My first and last emphases.)

The phrase 'its most expendable members' is not quantificationally precise. For the claim about *least* damage to be true, the precise quantifier should be 'exactly one of its most expendable members'. But what does Fuhrmann do via his technical definition? He changes the quantifier to '*all* of its most expandable members'! Fuhrmann's so-called 'minimal' contraction therefore involves getting rid of *all* $<$-minimal elements. Let us call this the *en bloc* reading. To opt for the *en bloc* reading of the phrase 'its most expendable members' is a gaffe, given Fuhrmann's opening description of the project as that of describing 'how theories ought to *minimally* change so as to incorporate new information or retract old, superseded, information'.

Another blunder lurks in the motivation Fuhrmann offers for what he calls 'mind-opening' contractions. The error reveals a deficient understanding of perfect validity—or, in Fuhrmann's terms, of entailment sets. He writes (*loc. cit.*, p. 192)

> ... we do not want A to be the only minimal (under $<$) member of some entailment set for $\sim$A in the A-extended base. For then the A-free part of such an entailment set may survive the contraction by $\sim$A only to give us back $\sim$A after A has been added again.

But let us reflect for a moment on his envisaged possibility. Let Y be an entailment set for $\sim$A, and let X be its 'A-free part'. We are asked to imagine that $X, A \vdash \sim A$ (for A, note, is supposedly in Y). But any sane logic, relevant logic included, includes the rules of *reductio ad absurdum* (negation introduction); so this would mean that $X \vdash \sim A$. But X is a *proper* subset of Y (since it does not contain A)! Hence Y cannot, after all, have been an entailment set for $\sim$A. One doesn't mind having one's mind broadened, but to ask one to *open* it in this way is to ask a little too much.

[22] As Fuhrmann in effect does.

[23] Note that for Fuhrmann it would only be the Upward pass that came into play. He makes no provision for the Downward pass required on the new perspective. And if in reply he were to claim that the effects of Downward passing would anyway be secured somehow 'in the wash' by operating on *bases* of theories rather than on the whole theories themselves, one could point out that he is thereby limiting himself to the foundationalist case. The Staining Algorithm, by contrast, applies to both the foundationalist and the coherentist case.

to capture the requirements of minimum mutilation. The 'single <-minimal sentence at a time' approach, by contrast, is much more likely to yield an understanding of what sorts of heuristics will reliably produce minimally mutilated results in the shortest possible time (compatibly, of course, with their not being flagrantly counterintuitive on the new perspective).

It is worth enquiring after the reasons why someone might favour the *en bloc* approach over that of the 'single <-minimal sentence at a time' approach. Perhaps it has something to do with the feeling that if one is to give up any particular sentence from among several candidates that are all <-minimal within a given premiss set, then one ought to have a differentiating reason for choosing the sentence that one does. Perhaps, the suggestion goes, we operate on a *principle of sufficient reason* in combination with the principle of minimum mutilation.[24] And, the thought would be, if we had such a differentiating reason it would surely have been registered in the entrenchment relation, so that the sentence in question would have been less entrenched than its stablemates in that premiss set, and not tied with them for <-minimality.

Here I want to make a bold suggestion. Strange though it may seem, the principle of sufficient reason will really only make itself felt if it is made to operate via the deliverances of an *holistic* contraction process strongly conditioned by the principle of minimum mutilation and based on the 'one <-minimal sentence at a time' approach. If we run through the many alternative routes of contraction that arise from making the 'arbitrary' single choices enjoined by minimum mutilation, we may find that *holistically*, when particular sequences of choices are made, we find *all sorts of reverberating and connected* reasons for retaining this particular premiss (in a perfect proof of some unwanted conclusion) in question, rather than those other ones. What may look, locally, like a choice (of premiss to stain) made with 'insufficient reason' may, on the contrary, be a choice endowed with *more than sufficient* reason when one pursues, holistically (in the way prescribed by the Staining Algorithm), all the knock-on effects, within our belief scheme, of the choice in question. It may turn out that the reason that emerges for dropping this one rather than those other ones is not obvious, and can only be brought out by comparing the results of different contractions. Moreover, such 'holistic' reason is not one that we can expect to find registered, in advance, in the 'more obvious' facts of entrenchment, for entrenchment might be a more local matter. At least, a theorist with a rather sparse entrenchment relation cannot be convicted for not having already extended the relation in the dim light of these distant ramifications. Thus, ironically, the *en bloc* approach, appealing as it does to the principle

[24] Here I am endebted to Alex Oliver.

of sufficient reason, obliterates the possibility of generating materials on which the principle of sufficient reason could then manifestly operate!

I think it is their failure to contemplate this delicately textured manifold of possibilities that has led various writers[25] to opt for the *en bloc* approach rather than the 'single <-minimal sentence at a time' approach. The *en bloc* approach may yield a certain tractable mathematical definiteness to the operation of contraction, by guaranteeing the uniqueness of its result; but the contractions enjoined by the *en bloc* approach are, to my mind, methodologically unsound. The mathematical representation theorems that it enables, no matter how elegant they may be, lose their interest because they are based on *misrepresentations* of the manner in which, in general, one ought to contract a theory! By contrast, the 'single sentence at a time' approach is normatively correct. Moreover, it both motivates, and promises to exploit the results of, a computational or algorithmic approach.

Given (i), (ii), and (iii) above, I would suggest the following procedure as globally sensible. But it should be stressed that any failure or shortcoming of the suggested procedure should not be taken as reason to abandon altogether the approach via (i), (ii), and (iii); it would be reason only to look for a finer-tuned alternative than they would thereby have been shown to constitute:[26]

> Consider the union of those Δ_i that are free of stain. Choose some least entrenched member F_0 of this union, that also occurs in as many as possible of the as yet unstained Δ_i. Stain F_0. Now consider the union of those Δ_i still free of stain. Choose some least entrenched member F_1 of this union, that also occurs in as many as possible of the as yet unstained Δ_i. Stain F_1.... Repeat this process until there are no Δ_i free of stain.

In what order should upward and downward staining be performed? It might seem from our discussion of the toy example earlier that we would need to apply downward staining first, in order to get the desired result [~C] when contracting [B,~C] with respect to B. It might appear that if we were to apply upward staining first, we would get the wrong result [B ∨ C]. But this would be a mistake, albeit an understandable one on the part of one who persisted in construing a theory merely as a set of sentences closed under logical consequence. As it happens, because we conceive of a theory as a set of entries carrying justifications for sentences, the order in which Upward and Downward staining is carried out is irrelevant. In the example of the theory [B,~C], the proof of the inference

[25] Such as Alchourrón and Makinson [1985] and Fuhrmann [1991].
[26] Here I am endebted to Jeremy Butterfield.

$$\frac{B \vee C \quad \sim C}{B}$$

is not one of the pedigrees for B. Hence even with Upward staining from B we shall not be considering which of $B \vee C$ and $\sim C$ to get rid of. Therefore we cannot even get ourselves into the mistaken position of holding $\sim C$ to be the most vulnerable culprit.

The two steps Upwards and Downwards can therefore be carried out independently of each other, or in parallel. Each will be applied until it can be applied no further: that is, until one runs out of applicable instances for each step. Then all the entries for sentences stained by that stage will be gathered together. If all pedigrees are from ultimate justifiers, we can halt at this stage and jettison all the entries with stained sentences. But if the pedigrees involve premisses that themselves in turn require justification, the process will have to be repeated as often as is necessary to climb all the way up the branches of the justificatory tree.

The staining algorithm offers a different approach, in response to the problems inherent on the limited perspective, than those offered by other writers who have shown some awareness of these problems, but then resorted to contractions of bases for its solution. The latter approach cannot deal with theories equipped with coherentist justificatory structure. The staining algorithm can.

8 Representing theories and computing contractions

At p. 36 of *Knowledge in Flux* [1988], Gärdenfors writes

> it seems to be a matter of fact that people do not keep track of the justifications for their beliefs. The main reason for this is that it would soon lead to a combinatorial explosion, and it is a matter of the *economy of thought* to avoid cluttering one's mind.

But there is a trade-off between the resources needed to represent a theory and the resources needed to effect a revision. So the matter is not quite so simple as Gärdenfors would have us believe. On the contrary, I think Gärdenfors is giving hostage to empirical fortune with this claim, and not fully appreciating the trade-offs, well known to computationalists, between economy of representation and economy of processing. It would be advisable, to say the least, to avoid any a priori dogmatism on this score until we know a little more about the computational neurobiology involved. Gärdenfors cites Levi[27] in support of a coherentist view

[27] See Gärdenfors [1990], p. 30.

(according to which, he mistakenly thinks, one would have no interest in justificatory pedigrees):[28]

> whether pedigree is traced to origins or fundamental reasons, centuries of criticism suggest that our beliefs are born on the wrong side of the blanket. There are no immaculate preconceptions.

But this doesn't support a coherentist view over a foundationalist one, regarding the *structure of internal justifications* within our belief systems. All it does is question the external security of whatever foundation a foundationalist system happens to have. In no way does this entail that the *structure* of justificatory pedigrees will not be a well-founded one.

Representing a theory by means of entries with ultimate justifiers makes great demands on memory (regarding the sequences of justifiers), but means also that the contraction algorithm makes only one pass in order to reach a result. The demands on memory do not make the richer representation 'unfeasible'. For feasibility—which is usually understood as computability within polynomial time—is to be judged by the time taken to produce an output, *given* an input. It does not matter how long the inputs are, provided they are finite. If there is a polynomial function f such that for all n, no matter how large, the time taken to compute a result on any input of length n is bounded above by f(n), then the computation is feasible, no matter how long it might take to feed in the (finite) input of length n! So merely making our inputs longer and more complicated has no implications at all for the feasibility of the computations that are to be performed with them.

It is no objection, either, to the representation-of-justifiers approach to cite empirical evidence, as Gärdenfors does,[29] about the outcome of psycho-epistemic experiments:[30]

> beliefs can survive potent logical or empirical challenges. They can survive and even be bolstered by evidence that most uncommitted observers would agree logically demands some weakening of such beliefs. They can even survive the total destruction of their original evidential basis.

Such hand-wringing obeisance to the data about human stupidity is entirely beside the point. The aim in any branch of AI that is worth the title is surely to achieve prosthetic extension of the range of *ideal human competence*, not a computer-reinforced farrago of fallacy and shoddy thinking. God forbid that the AI community should ever deliver to us the Cray version of the silly people studied in the experiments by Ross and Anderson, for service in the laboratory or in the study. The more indis-

[28] Quote taken from Levi [1980], p. 1.
[29] Gärdenfors [1990], p. 31.
[30] Quote from Ross and Anderson [1982], p. 149.

putable and repeatable the findings reported by Ross and Anderson turn out to be, the more urgent, I contend, is the need for a normative model of theory change such as the one that I am proposing.

9 The normative nature of the enterprise of knowledge

The point to be stressed here is that the logic of theory change is as normative an enterprise as the internal logic behind theory closure. Just as the latter is concerned with such questions as:

> How *ought* we to represent the structure of thought and language?
> How *ought* we to use those structures in inference?
> What are the *norms* of reasoning within such-and-such a language?

so too is the former concerned with such questions as:

> How *ought* a rational agent, given such-and-such a system of beliefs with such-and-such justificatory pedigrees already to hand, and with such-and-such a pattern of epistemic entrenchment, go about *dropping* one of her beliefs so as to retain as much as possible of what she formerly believed, but to retain it only with such justifications as are *entitled*, genuinely, to survive in that process?

Furthermore, *pace* Gärdenfors, I cannot agree that[31]

> A principle of intellectual economy would entail that it *is* rational to neglect the pedigrees of one's beliefs. To be sure, we will sometimes hold on to unjustified beliefs, but the erroneous decisions caused by this negligence have to be weighed against the cost of remembering all reasons for one's beliefs.

I would like to believe that Gärdenfors must have neglected the pedigree of this particular belief; for that would at least entail that it once had one! Where does such a breath-taking claim get its credentials? Introspection? Cognitive psychology? Computational complexity theory? Neural network theory? Neurobiology? Folklore? Why is it that folk are usually able to tell one pretty convincingly and swiftly, on reflection, why they hold a particular belief of theirs when one points it out to them and questions it? How is it that scientists are able to set about rational revisions at all, if not by being able to recall, and retrace, and reassess, the links of justifications in the pedigrees they have for their various beliefs, individually and as a community?

Gärdenfors gives short shrift to the AI community's well-known truth-maintenance systems because they are allegedly 'inefficient'. But nothing can be concluded from this accusation at this stage. The alleged 'inefficiency' could be the fault of bad implementation, not the fault of the intrinsically

[31] Gärdenfors [1990], p. 31 *infra*.

available algorithms. Or it may be the fault of using silicon-based machines whose architecture is not neurobiologically faithful.

One is prompted to ask why Gärdenfors moves so easily from an assumption about 'cost' (nowhere specified) in maintaining a certain kind of belief system, to a conclusion about what kind of belief system, and what kind of materials on which to base our methods of contraction and revision, it would be *rational* to have.[32] Who ever maintained that ideal rationality itself was the outcome of corner-cutting or of satisficing, or of quick-and-dirty methods that work 'almost always' but have disastrous consequences now and again? Isn't it rather that we have an ideal of what would be rational, which may be extraordinarily demanding, and that it is *to the extent that one approximates* that ideal by means of any of these latter methods that their own utility or serviceability is to be judged?

If so, why shrink then from the task of specifying exactly what the ideal belief-structure of a rational agent would, in general, have to incorporate? If such an ideal indeed involves representing each belief with its whole pedigree, or by means of some method whereby such pedigrees can be reconstructed, then so be it. *Normative* modelling, after all, is the name of the game. Let us get the *normative* model in place, and then test it for computational feasibility.

Again, the comparison with ordinary logic is a good guide. We can be grateful that Frege, and other great logicians since, did not proceed by asking the person in the street for their responses to various proposed forms of inference in first-order logic. The literature on the psychology of deduction, and the experience of teaching introductory logic to any class of undergraduates, makes that painfully clear. Rather, Frege and others set out a system of *norms* about whose *implementability* we were subsequently able to enquire.

We now know a great deal about the limits, in principle, to any cybernetic embodiment of 'full competence' with such systems of norms. This hasn't prevented automated deduction, however, from being a fruitful and exciting discipline. It has, rather, sharpened the problematic: given Church's undecidability theorem, and given the basic belief that we are probabilistic neural networks, how on earth do we manage, in logic and mathematics, what we do?[33] In like fashion, we could have a normative

[32] I am adverting here, of course, to the *kind of structure* the entries themselves might exhibit, and the kinds of relations, justificatory or otherwise, that these structured entries might be represented as bearing to one another. I am not talking about the *contents* of any particular beliefs that it would be rational for us to hold.

[33] Penrose, in his book *The Emperor's New Mind* [1990] of course does not share the basic belief mentioned. His own reasons for thinking that we are not automata strike me as confused. That, however, is another story.

logic of theory change as an imperfectly attained ideal, and investigate just what aspects of it we ourselves manage to 'implement' in our own practice of changing our theories.

There is, of course, a middle ground in the matter of representation of beliefs with their justificatory pedigrees. We could try representing a theory instead more economically by means of entries with justifiers invoking as premisses previously justified claims (but without *their* justifications). This would make weaker demands on memory (for the justifiers can be shorter), but it would entail correspondingly greater processing effort as we repeat passes of the Upwards and Downwards steps to make sure that the stain has reached all the sentences that it should. Logical cut can reduce the costs of representation and of proof discovery; but it can increase the costs of contraction. Perhaps that is why, in a field like pure mathematics where the prospects for contraction are so negligible, logical cut is so important. And perhaps we could also expect that with theories involving more vacillation, more tentative conjecturing, and subject to more frequent demands for contraction, we would find it both rational and cost-effective to keep a slightly more explicit representation of the ultimate pedigree of any of the claims to which we make passing commitment.[34]

Moreover, there is a further reason why with inductive warrants we would be loathe to provide justifications broken down into stages to be telescoped with applications of cut so as to yield the 'ultimate' grounding of an inductive conclusion. This is that strong inductive support is not transitive. It is not even transitive on consistent sets of premisses. Since inductive inference from an inconsistent set is impossible anyway, one would only ever contemplate such applications of cut as produced consistent sets of premisses overall. But even in such cases, strong inductive support, unlike deducibility, is not transitive.

10 The finite and the infinite, and the working identity of a theory

In the finite case our 'algorithm' is reminiscent of the algorithm for finding the R-closure generated by an element S in the field of a two-place relation R. The relation R in the case of theory contraction happens to be dynamic, in the sense that as the Upwards and Downwards steps are performed the

[34] In arguing against representing beliefs with their justificatory pedigrees, Gärdenfors ([1990], pp. 31–2) continues after my last quotation above from p. 31 *infra*, 'The balance will certainly be in favour o[f] forgetting reasons. After all, it is not very often that a justification for a belief is actually withdrawn and, as long as we do not introduce new beliefs without justification, the vast majority of our beliefs will hence remain justified.' If he really believes that, one wonders what prompted his theoretical interest and work in the problems of theory contraction and theory revision in the first place!

particular new elements chosen for staining (*i.e.* for inclusion in this choice-dependent R-closure) serve to determine a range of possibilities for where the relation R might hold for the next stage. At each stage one chooses some, but not necessarily all, the (unstained) R-successors of elements most recently chosen for staining. In the finite case the 'closure' can be completed in polynomial time.

Because at each stage we choose *some*, but not necessarily *all*, the (unstained) R-successors of elements most recently chosen for staining, there opens up the possibility of considerable variation among alternative outcomes to the contraction process when we prompt for alternative solutions and it backtracks.[35]

Our 'algorithm', note, is not in general finite. For the Upwards trawl could (at least after the initial stage) catch infinitely many items $(S,\langle\Pi\rangle)$, and any of those items could have infinitely long proof sequences $\langle\Pi\rangle$; and the Downwards trawl could catch infinitely many items $(S,\langle\Pi\rangle)$ such that $\langle\Pi\rangle$ is stained. But we can make sense of a *stage* in the application of the combined Upwards–Downwards procedure. And despite the possibility of infinite furcations just adverted to, the procedure could halt at some finite stage. If not, then we shall need to define the set of stained sentences in the obvious way at limit stages, and provide for continuation of the process into the transfinite. We know that there will eventually come a stage at which no more sentences can be stained. Let us call this the *closure ordinal* for the sequence of choices that has led to it.

Even in the infinite case I believe our *method* of contraction, proceeding as it does along the lines of justification delineated *within* the theory concerned, is a sounder and more realistic account of what ought to go on, and of what probably does go on, in the mind of a competent theorizer. It is a very 'predicative'[36] procedure, unlike those of Gärdenfors *et al.* These other writers characteristically define their contraction functions by appeal to large classes of subsets of the theory being contracted, subsets that are maximal in respect of failing to imply the unwanted sentence, and among which choices are to be made by appeal to global facts about entrenchment. It is an impredicatively 'top–down' method, one which would be unnecessarily cumbersome in the finite case, and which would not lend itself to good finite approximations in the infinite case. Our method, by contrast, minimizes the computational work in the finite case, and would yield, I think, an optimal answer in the infinite case, if

[35] I have in mind here a depth-first algorithm, as it would be if implemented in Prolog.

[36] I use this term metaphorically; anyone with constructivist sensibilities will know what I mean. I am adverting to the need to compute larger objects from 'within', using material in the form of mathematical atoms and relations 'bottom–up'. The 'top–down' methods, by contrast, involve procedures like the intersection of families of sets, each of them impossible to compute from the materials given.

one could imagine it being carried out on potentially infinite sets of pedigrees of sentences in the theory.

Our method would also work well for *finite partial developments* of axiomatized theories. Any axiomatized theory based on an effective set of axioms has both an *ideal* identity and, at any stage of its development, a *working* identity. The ideal identity is given by the infinite closure of its axioms under whatever relations of deducibility and/or inductive support are relevant. (We may also add: along with their pedigrees—all possible pedigrees countenanced by the relations of deducibility and/or inductive support.) The working identity is given by the finitely many justifications that have thus far been provided for claims recognized to be part of the theory. Naturally these include the axioms to begin with; and, assuming perfect communal memory and intellectual apprenticeship, and no misgivings leading to contractions, the set of pedigreed claims would grow monotonically with time. The claims would not necessarily come tagged with all possible pedigrees as they would in the ideal case. At any working stage, only finitely many different pedigrees would have been discovered for any given claim. But this means we can also picture theoretical progress rather more realistically along two dimensions: that of adding an entirely new claim, along with its first ever pedigree; and that of consolidating an earlier pedigreed claim by supplying yet another pedigree for it.

This provides a rather more realistic model of the problem of contraction that would pose itself to any scientist working with the theory. She would have the axioms, and finitely many claims supported by those axioms. Moreover for each such claim she would have finitely many pedigrees establishing it within the theory. I submit that the problem of contraction can be solved only *modulo* what I have called the working identity of the theory concerned. Provided only that the scientist applies the procedure of staining correctly to the theory at its current stage of maturity, she will have met the demands of rationality that it is licit to pose.

Naturally one can be in what might be called 'the finite predicament' when using the working identity of the theory. Because the working identity of the theory is but a finite approximation to its ideal identity, even a correct application of the method of staining might produce non-ideal results. I said earlier that one should look down from S as a *justifying premiss* to conclusions R in T that rest essentially on S and other exiles, in the sense that every proof of R in T draws on S or one of those other exiles as a premiss.

Now in the finite predicament this will mean 'every proof of R *so far furnished* in T draws on S or one of those other exiles as a premiss'. And it may be compatible with this being the case that, ideally, there is some (as yet undiscovered) proof in T of R that does not draw on S or any of those

other exiles as a premiss. In such a case the procedure will have been in error in getting rid of R along with S and those other exiles. But there might nevertheless be the following consolation: if, ideally, R should have survived, then (assuming there are no further contractions) it might still be possible for that as-yet-undiscovered justification for R to be produced at some later working stage of the contracted theory. The theory T would, as it were, merely have had a haircut that did not succeed in changing its hairline. Whether this consolation can be guaranteed in all cases, however, is a delicate matter. For the rejection of R might have precipitated further rejections of other sentences that would have been needed for that as-yet-undiscovered justification of R in the ideal theory before the contraction.

If we are being computationally realistic, however, such a 'mistake' may just have to be tolerated stoically. When a model of godly competence appeals to infinitary syntactic objects (such as the logical closure of a set of sentences, each with its potentially infinite set of justifications from the axioms), then a parallel model of our finitary handling of perforce finite fragments of such infinitary objects will have to put up with some occasional rough in order to be generally smooth.

This indeed is a problem that is simply skirted over by Gärdenfors *et al.* To be sure, they avoid the potential embarrassment just pointed out with downwards staining in the finite predicament, by ignoring downward staining altogether. But by the same token they are unrelentingly *ideal* and *infinitary* in their treatment of what in my terminology would be called upward staining. For their 'construction' of the contracted theory proceeds in the following ideal and infinitary fashion. First, the theory T is taken to be logically closed already, before one contracts with respect to S. They consider all maximal subsets of T not implying S, and choose one such subset on the basis of considerations concerning entrenchment.[37] This is all very well, and roughly corresponds to our treatment of upward staining in the ideal and infinitary case.

11 On implementation

It is not clear, however, how AGM might set about implementing such a method at a finite working stage. Since one of the claims for the AGM theory[38] is that it will be of interest to those working in AI, it is fair to say

[37] This holds also for 'relational partial meet contraction', where the relation among sets is a form of entrenchment relation, albeit in a generalized sense.

[38] See the preface in, and publisher's blurb for, Gärdenfors [1988]. Moreover, Gärdenfors [1990] begins with the claim that 'solutions to these problems will be crucial for any attempt to use computers to handle *changes* of knowledge systems. Problems concerning knowledge representation and the updating of such representations have become the focus of much recent research in artificial intelligence (AI).'

that the criteria by which the AGM theory should be assessed alongside its rivals should at least include some consideration of their respective merits and demerits when it comes to implementing them. That is, one would like to have some idea of how good an *automated theory contractor* one could program on the basis of the theoretical ideas and methods on offer.

Having no implementation is arguably worse than having an implementation that suffers from the finite predicament. Finite beings simply have to do the best that they reasonably can. Is there, in the finite predicament, a potential for mistakes with our step of upward staining, parallel to that described above for downward staining? It would appear that there is. For example, suppose we are contracting the theory at some finite working stage with respect to the sentence S. We consider the set of justifications for S *at that finite working stage*. We stain some least entrenched premiss within each so far unstained justification for S. But might there not be some as yet undiscovered justification for S, one of whose premisses should also be stained in this way, but which will escape staining simply because the justification in question is not within purview at this finite stage? The answer has to be an honest affirmative. Until such time as the contraction with respect to S might be countermanded, the injunction 'Do not commit yourself to S!' will have to be borne in mind.[39] If the as-yet-undiscovered justification for S is stumbled upon at some future working stage of the supposedly S-contracted theory, then this should set in train once more a process of upward and downward staining, starting with the rogue intruder S.

Not being logically omniscient, the best we can do at any stage is to get rid of all the obvious and known ways to justify S. We cannot in general guarantee that the logical back door is forever closed to S. The price of freedom from explicit justification for S is eternal vigilance.

In this way the situation is not unlike that of ensuring consistency. We are under the standing injunction 'Be consistent!'—that is, 'Do not commit yourself to absurdity!' So the picture on offer of a working theory is that it will consist of a core axiomatization which is hypothetical in two senses.

First, there is the usual sense according to which it will contain hypotheses designed to predict and explain various assertions present with their own face value, so to speak (protocol sentences, or sentences on Quine's periphery, like his pegged observation conditionals). Such hypotheses are, as Popper has made us aware, in the theory tentatively and speculatively as conjectures, their logical power making them at once useful and vulnerable to further testing.

[39] This would be like putting S on the 'outlist' of Doyle's truth maintenance system. See Doyle [1979].

Second, there is a sense in which the axiomatization of the theory is hypothetically assumed to satisfy a number of injunctions (such as 'Do not commit yourself to absurdity!', 'Do not commit yourself to S!' ...) which accumulate over the course of one's research. The theory is liable, at any working stage, to be proved to fall foul of one of these injunctions. At the very worst, an inconsistency will be discovered; less dramatically, it may be discovered that there is a justifying route within the theory after all to one of those sentences S to which we have been enjoined not to commit ourselves, and of which we had thought (mistakenly, we now realize) that the theory had been purged.

I see no reason why, once we acknowledge the finitary predicament with regard to the consistency problem, we should not acknowledge also that the finitary predicament will affect us for *all other sentences to which we might wish to avoid committing ourselves*. Especially when our logic of justification is effectively undecidable, and even when it is decidable but not very tractable, this danger will lurk; and we shall be forced to acknowledge that any working stage of our theory is hypothetical in the second sense as well.

We have been considering contraction as a process of removing offending items by looking *downwards* as well as *upwards* along lines of logical propagation. The method adopted by Gärdenfors, Makinson, and Alchourrón was to look only upwards and attend to entrenchment. Moreover (in our terminology), they stain at least one least entrenched premiss[40] of any proof of a stained sentence S. (This requirement can actually be in tension with that of minimum mutilation.) But they neglect altogether the need to look *downwards* as well! The resulting difference between their method (displaying the limited perspective) and my method of upward *and downward* staining (displaying the new perspective) is brought out vividly in the very simple example already discussed above: that of contracting the theory [B,$\sim$C] (where B $>$ $\sim$C) with respect to B.

12 Further reflections: choice of logic, and inductive support

It is worth noting that our method of upward and downward staining is invariant with respect to choice of logic. *Any* deducibility relation generated by proofs of conclusions from premisses will do. The notion of perfect proof is invariant across different deducibility relations. (A proof of A from the set X of premisses is perfect just in case A cannot be proved from any proper subset of X.) It is worth making this point about invariance with respect to choice of logic clear for two reasons. First, certain results of the AGM

[40] Indeed, they are not even globally sensitive in their choice of premisses to be stained. And some of their contraction methods involve getting rid of *all* least entrenched premisses, rather than choosing only one of them.

theory depend, for their proofs, on the fact that the underlying logic for the closure of the theories is classical. Secondly, the reader might form the impression, given the mention of perfect proof, that the treatment here recommended would work only if the underlying logic were a relevant logic (for that is the usual context in which logicians get interested in the notion of perfect proof). But of course this is not so. Indeed, the treatment is so general that it would work not only for the standard deviant logics (minimal, intuitionistic, Anderson-Belnap relevance logics ...) that have unrestrictedly transitive deducibility relations, but also for logics whose deducibility relations are not unrestrictedly transitive.[41]

Indeed, even restriction to *deducibility* rather than *inducibility* is unnecessary. With upward staining, the proofs whose premisses are to be stained do not have to be *deductive* proofs. It might be thought that only a deductive proof could *force* a conclusion upon one if one accepted the premisses of the proof. But if those premisses had originally provided inductive justification for the sentence to be eliminated, they could continue to do so even after a contraction operation that left them in the contracted theory. So we could have just as much (inductive) reason *after* the contraction as we did before it, to believe the sentence so justified. So with upward staining we would do well to consider relations of inductive support as well as the logical deducibility relation. Likewise with downward staining, it would be wise to stain any conclusion even of *inductive* proofs or arguments that can be supported only by such arguments as have a stained premiss. A conclusion arrived at inductively, and which depends essentially for its support on propositions that are to be thrown out, should be thrown out with them. Thus the new perspective encompasses relations of inductive support as well as the deducibility relation of the logic of the language concerned. This is another respect in which it is more general than any of its competitors in the AGM tradition. For they all work with consequence relations satisfying the structural conditions of transitivity and dilution—which rules out the relation of inductive support.

13 On making choices

The only complication attending our method arises from the possibility of distinct choices of least entrenched premisses for staining. Inspection of our 'algorithm' reveals that each such choice precipitates a particular collection of items at the next stage. The membership of this collection

[41] For an account of these systems and their importance, see Tennant [forthcoming]. I see no obstacle, in principle, to generalizing the treatment so as to deal also with non-monotonic relations of inferential support. But the further subtleties involved would lie beyond the scope of the present paper.

depends on the least entrenched premiss chosen. It determines, in its turn, a choice of further sentences for staining; and determines also a closure ordinal. So in general each particular sequence of choices that are made as genuine alternatives posed themselves will have its own peculiar outcome, with an associated closure ordinal. Each such outcome will be a contracted theory all right, which fails (in the ideal case) to imply the original sentence S with respect to which one was contracting. And each such outcome will (in the ideal case) avoid having in it sentences that would have depended, for their presence, on ones that had to be thrown out. But now we have to ask: will there be, among all such outcomes, some one (or few) to be favoured on grounds of minimum mutilation, or of speedy completion of the contraction process, or both?

I conjecture that the principle of minimum mutilation and the requirement that the closure ordinal be as small as possible will probably go hand-in-hand. Indeed, any two results that enjoy the minimum closure ordinal will not be discriminable on grounds of minimum mutilation alone. Some further criterion for unique choice would have to be applied. The problem of multiple results could be avoided altogether, of course, simply by staining *all* least entrenched premisses of any unstained proof of a stained sentence.[42] But as pointed out above, this would in a certain sense *maximize* the minimum mutilation being inflicted on the theory! It turns a Sophie's choice into a Hobson's choice. Mathematically, it brings us uniqueness, but at an enormous price. It would also involve more computational work than would a quick-and-dirty method that sought (on an arbitrary list-ordering) the *first* least entrenched premiss, stained it, and moved on to the next proof to be disabled. Minimum mutilation, even if it sometimes involves quirky favouritism, is at least an *efficient* requirement. And there is a potentially richer set of possible outcomes to study with the aid of computer implementations, if one systematically varies the single choices to be made from among the alternatives when applying the staining algorithm. The Hobson's choice method blurs that structure entirely, and might thereby miss the opportunity to discover interesting regularities and singularities in the process of theory change.

In actual cases, of course, even the most rational of human beings do not deal with infinite theories, nor with infinitely many ways of contracting a given theory with respect to a particular sentence. They do not bear in mind the infinitely many consequences there will be from $\vee$-introductions and the like. They work instead with finite bases or axiom sets, and provide for each sentence in their theory at most finitely many pedigrees of justification. If we limit ourselves, therefore, to finitely axiomatized theories (or at least

[42] Such, as noted above, is the strategy favoured by Fuhrmann [1991].

theories with finitely many axiom schemata), whose items $(S,\langle\Pi\rangle)$ are finite, then contraction by our recipe will be fully algorithmic.

In the finite case there is another advantage worth noting if one's concern is to provide a normative account of how theories ought to be contracted. We have been treating statements of a theory as items $(S,\langle\Pi\rangle)$ carrying a sentence S and a sequence $\langle\Pi\rangle$ of proofs of that sentence within the theory. A variation on this is to put in place of the proofs themselves just their premiss sets. Either way, the idea is clear: the extra material provided in this way gives us good leads in working out what else has to go when we try to get rid of one particular statement of the theory. Now on a normative account, the sequence $\langle\Pi\rangle$ ought to contain *all possible* justifications of S within the theory. It would not do to leave any possible proof Π (or, more precisely, premiss set) out of account. This is because the premisses of such Π might survive the method of staining. The sentence S would thereby survive as a potential conclusion of this as-yet-undiscovered proof Π. This would mean that the contraction had failed to get rid of S after all. The point in favour of the finite case is this: such a possibility would be reduced to a minimum, and, especially in the logically decidable case, would be in principle preventable.

The requirement to look downwards when applying the recipe for contraction may be honoured in the breach. One of our epistemic failings is to forget the original justifications we may have had for coming to believe a proposition. This can hold even for propositions that depend essentially (still) on some other propositions that are now to be discarded. In that case the Downwards step is not carried out fully, and the tainted progeny survives without stain. This is simply another example of a competence-performance gap, like any other fallacy in reasoning. It arises through the loss of deductive information as to the provenance of a claim with which we have become familiar.[43] The process may even be psychologically plausible. Items $(S,\langle\Pi\rangle)$ may well be stored by being indexed primarily with the sentence S and only secondarily with the various members of $\langle\Pi\rangle$ that justify it. The primary indexing may be more robust or long-term than the secondary. Hence, if the traces of $\langle\Pi\rangle$ degrade, one is left with $(S, \ldots)$. This might receive default treatment as $(S,\langle S\rangle)$, that is, as a self-justifying claim rather than one with a non-trivial pedigree descending from other more basic claims.[44]

[43] Here we might get a model of *performance* to go with our model of competence, and which might account for the observations reported above from Ross and Anderson [1982].

[44] By contrast, a well-trained mathematician or forensic scientist or trial lawyer is much less likely to let the various $\langle\Pi\rangle$ degrade; for these are people whose professional occupations depend on their ability to remember chains of justification, purported and actual, and to appraise them and to reproduce them on demand.

Indeed, the limited perspective will rearise simply by treating every item as self-justifying! In that case the step we call Downwards will garner nothing, and all the contractual work will be done by the step called Upwards. This step will not, to be sure, be able to be carried through just by attending to the information internal to each entry. (Rather, one will have to ask afresh: what proofs might there be of S within the theory if we ceased to regard each claim as self-justifying, and bore in mind that we are committed to the consequences of the claims that we accept?) But as we have seen, there will be an epistemic price to pay for this neglect or forgetfulness with regard to justificatory pedigrees: namely, we shall get *wrong answers* by intuitive standards that are eminently clear in the examples given above.[45]

Department of Philosophy
The Ohio State University and
Churchill College, Cambridge

References

Alchourrón, C., Gärdenfors, P. and Makinson, D. [1985]: 'On the Logic of Theory Change: Partial Meet Functions for Contraction and Revision', *Journal of Symbolic Logic*, **50**, pp. 510–30.

Alchourrón, C. and Makinson, D. [1982]: 'On the Logic of Theory Change: Contraction Functions and their Associated Revision Functions', *Theoria*, **1**, pp. 14–37.

Alchourrón, C. and Makinson, D. [1985]: 'On the Logic of Theory Change: Safe Contraction', *Studia Logica*, **44**, pp. 405–22.

de Kleer, J. [1986]: 'An Assumption-based TMS', *Artificial Intelligence*, **28**, pp. 127–62.

Doyle, J. [1979]: 'A Truth Maintenance System', *Artificial Intelligence*, **12**, pp. 231–72.

Fuhrmann, A. [1991]: 'Theory Contraction Through Base Contraction', *Journal of Philosophical Logic*, **20**, pp. 175–203.

Fuhrmann, A. and Morreau, M. (*eds.*) [1991]: *The Logic of Theory Change*. Lecture Notes in Artificial Intelligence 465, Springer.

Gärdenfors, P. [1982]: 'Rules for Rational Changes of Belief', in T. Pauli (*ed.*), *Philosophical Essays Dedicated to Lennart Aqvist on his Fiftieth Birthday*. Philosophical Society and the Department of Philosophy, University of Uppsala.

[45] I am grateful to members of the audiences for valuable discussion when ancestors of parts of this paper were read to the Moral Sciences Club in the University of Cambridge, to the Department of Logic and Metaphysics in the University of St. Andrews, and to the Department of Philosophy in the University of Edinburgh. I wish also to thank David Papineau for his valuable editorial suggestions, which have led to significant improvements in exposition.

Gärdenfors, P. [1984]: 'Epistemic Importance and Minimal Changes of Belief', *Australasian Journal of Philosophy*, **62**, pp. 136–57.

Gärdenfors, P. [1988]: *Knowledge in Flux*. Cambridge, MA: MIT Press.

Gärdenfors, P. [1990]: 'The Dynamics of Belief Systems: Foundations vs Coherence', *Revue International de Philosophie*, **44**, pp. 24–46.

Gärdenfors, P. and Makinson, D. [1988]: 'Revisions of Knowledge Systems using Epistemic Entrenchment', in M. Y. Vardi (*ed.*), *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*. Los Altos, CA: Morgan Kaufmann Publishers.

Hansson, S. [1989]: 'New Operators for Theory Change', *Theoria*, **LV**, pp. 114–32.

Levi, I. [1977]: 'Subjunctives, Dispositions and Chances', *Synthese*, **34**, pp. 423–55.

Levi, I. [1980]: *The Enterprise of Knowledge*. Cambridge, MA: MIT Press.

Makinson, D. [1985]: 'How to Give it Up: A Survey of Some Formal Aspects of the Logic of Theory Change', *Synthese*, **62**, pp. 347–63.

Makinson, D. [1987]: 'On the Status of the Postulate of Recovery in the Logic of Theory Change', *Journal of Philosophical Logic*, **16**, pp. 383–94.

Martins, J. [1991]: 'Computational Issues in Belief Revision', in A. Fuhrmann and M. Morreau [1991].

Nebel, B. [1990]: *Reasoning and Revision in Hybrid Representation Systems*. Lecture Notes in Artificial Intelligence 422, Springer.

Penrose, R. [1990]: *The Emperor; New Mind*. London: Vintage.

Ross, L. and Anderson, C. A. [1982]: 'Shortcomings in the Attribution Process: On the Origins of and Maintenance of Erroneous Social Assessments', in D. Kahneman, P. Slovic, and A. Tversky (*eds.*), *Judgement under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Tennant, N. [1984]: 'Perfect Validity, Entailment and Paraconsistency', *Studia Logica*, **XLIII**, pp. 179–98.

Tennant, N. [forthcoming]: 'Transmission of Truth and Transitivity of Deduction', in D. Gabbay (*ed.*), *What is a Logical System?* Oxford: Oxford University Press.