PART I: FOUNDATIONS OF REASONING

Section 1: Some Philosophical Viewpoints

Chapter 1: Change in View: Principles of Reasoning

GILBERT HARMAN

BELIEF AND DEGREE OF BELIEF

Probabilistic Implication

We have a rule connecting implication and reasoning:

> *Principle of Immediate Implication* That *P* is immediately implied by things one believes can be a reason to believe *P*.

Is there also a weaker probabilistic version of this rule?

> *Hypothetical Principle of Immediate Probabilistic Implication* That *P* is obviously highly probable, given one's beliefs, can be a reason to believe *P*.

Suppose Mary purchases a ticket in the state lottery. Given her beliefs, it is obviously highly probable that her ticket will not be one of the winning tickets. Can she infer that her ticket will not win? Is she justified in believing her ticket is not one of the winning tickets?

Intuitions waver here. On the one hand, if Mary is justified in believing her ticket is not one of the winning tickets, how can she be justified in buying the ticket in the first place? Furthermore, it certainly seems wrong to say she can *know* that her ticket is not one of the winning tickets if it is really a fair lottery. On the other hand the probability that the ticket is not one of the winning tickets seems higher than the probability of other things we might easily say Mary knows. We ordinarily allow that Mary can come to know various things by reading about them in the newspaper, even though we are aware that newspapers sometimes get even important stories wrong.

This issue is one that I will return to several times, but I want to begin by considering a suggestion which I think is mistaken, namely, that the trouble here comes from not seeing that belief is a matter of degree.

All-or-Nothing Belief

I have been supposing that for the theory of reasoning, explicit belief is an all-or-nothing matter. I have assumed that, as far as principles of reasoning are concerned, one either believes something explicitly or one does not; in other words an appropriate "representation" is either in one's "memory" or not. The principles of reasoning are principles for modifying such all-or-nothing representations.

This is not to deny that in some ways belief is a matter of degree. For one thing implicit belief is certainly a matter of degree, since it is a matter of how easily and automatically one can infer something from what one believes explicitly. Furthermore, explicit belief is a matter of degree in the sense that one believes some things more strongly than others. Sometimes one is only somewhat inclined to believe something, sometimes one is not sure what to believe, sometimes one is inclined to disbelieve something, sometimes one is quite confident something is not so, and so forth.

Reproduced with permission from Harman, G. (1986) *Change in View* (chapters 3 and 4). Cambridge, MA: MIT.

How should we account for the varying strengths of explicit beliefs? I am inclined to suppose that these varying strengths are implicit in a system of beliefs one accepts in a yes/no fashion. My guess is that they are to be explained as a kind of epiphenomenon resulting from the operation of rules of revision. For example, it may be that *P* is believed more strongly than Q if it would be harder to stop believing *P* than to stop believing Q, perhaps because it would require more of a revision of one's view to stop believing *P* than to stop believing Q.

In contrast to this, it might be suggested that principles of reasoning *should* be rules for modifying explicit *degrees of belief.* In this view, an account of reasoning should be embedded in a theory of subjective probability, for example, as developed by Jeffrey (1983), not that Jeffrey himself accepts this particular suggestion. In fact, this suggestion cannot really be carried out. People do not normally associate with their beliefs degrees of confidence of a sort they can use in reasoning. It is too complicated for them to do so. Degrees of belief are and have to be implicit rather than explicit, except for a few special cases of beliefs that are explicitly beliefs about probabilities.

Let me say why this is so. To begin with, Kyburg (1961) observes that the Immediate Implication and Inconsistency Principles would not be right even as approximations if belief were a matter of degree.

> *Immediate Implication Principle* The fact that one's view immediately implies *P* can be a reason to accept *P.*
>
> *Immediate Inconsistency Principle* Immediate logical inconsistency in one's view can be a reason to modify one's view.

Propositions that are individually highly probable can have an immediate implication that is not. The fact that one assigns a high probability to *P* and also to *if P then* Q is not a sufficient reason to assign a high probability to Q. Each premise of a valid argument might be probable even though the conclusion is improbable. Since one might assign a high degree of belief to various propositions without being committed to assigning a high degree of belief to a logical consequence of these propositions, Kyburg argues that the Logical Implication Principle is mistaken.

Similarly, each of an inconsistent set of beliefs might be highly probable. To take Kyburg's lottery example, it might be that the proposition, "one of the *N* tickets in this lottery is the winning ticket" is highly probable, and so is each proposition of the form, "ticket *i* is not the winning ticket," for each *i* between 1 and *N*. So one might believe each of these propositions to a high degree while recognizing that they are jointly inconsistent. Kyburg argues there is nothing wrong with this, so the Logical Inconsistency Principle is mistaken.

It is not just that these principles have exceptions. We have seen that they are defeasible and hold only other things being equal. But if belief were always a matter of degree the principles would not even hold in this way as defeasible principles. They would not hold at all.

It would be odd for someone to take this seriously in a routine matter. It is contrary to the way we normally think. Imagine arguing with such a person. You get him to believe certain premises and to appreciate that they imply your conclusion, but he is not persuaded to believe this conclusion, saying that, although you have persuaded him to assign a high probability to each of your premises, that is not enough to show he should assign a high probability to the conclusion! This is not the way people usually respond to arguments.

Or consider the following attitude toward contradiction. As Jack asserts several things, you observe that he has contradicted himself. His response is that he sees nothing wrong, since all the things he has asserted are highly probable. This is comprehensible, but it is again different from the normal way of doing things.

A normal reaction to someone's refusal to accept the conclusion of a clearly valid argument after he says he has been persuaded to accept the premises, if he gives Kyburg's reason, is to suppose that he does not really accept the premises after all, but only believes of each that it is probable. Similarly, we suppose that a person who says at least one ticket will win and also says of each ticket that it will not win does not really believe of each ticket that it will not win but merely believes of each ticket that it is unlikely that that ticket will win. We do not ordinarily think of this as like the case in which an author believes each of the things he or she says in a book he or she has written and also believes that, given human fallibility, at least one of the things he or she has said in the book must be false. Such a person is justified in having inconsistent beliefs, but that does not show that the Recognized Inconsistency Principle is incorrect. It only shows that the principle is defeasible.

Of course, to say one normally thinks of belief in an all-or-nothing way is not to deny one sometimes has beliefs about probabilities. More important, one often manifests a varying degree of confidence in this or that proposition as revealed in one's willingness to *act*, for example, to bet. But this does not show one normally or usually assigns *explicit* levels of confidence or probability to one's beliefs. The degree of confidence one has might be merely implicit in one's system of beliefs. Subjective probability theory can give an account of one's dispositions without being an account of the psychological reality underlying those dispositions.

It might be said one *ought* to operate using explicit degrees of belief. This would imply one should make much more use of probability theory than one does.

Similarly, it might be said that one's goals should be treated as matters of degree. Since different prospects are more or less desirable, one ought to assign them different degrees of "subjective utility." In acting, one should act so as to maximize expected utility.

I argue [in Chapter 9 of *Change in View*] that this is not right. But even if it were right, such an appeal to probability theory would not eliminate the need for reasoning in the sense of change in view. One's subjective probability assignments would never be complete. They would often have to be extended. To some extent they could be extended by means of the Principle of Immediate Implication by considering the immediate implications of one's current probability assignments and by allowing for clutter avoidance and other relevant considerations. Furthermore, there would also often be cases in which current subjective probability assignments would have to be changed, for example because they were not consistent with each other. The Principle of Immediate Inconsistency then has a role to play. And there are other cases in which one will want to modify such assignments, for example, when one discovers that a current theory would explain old evidence one had not realized it would explain (Glymour 1980, chap. 3). And whatever principles are developed for changing all-or-nothing belief will apply to changing degrees of belief, treating these as all-or-nothing beliefs about probabilities.

Conditionalization

Some probability theorists appear to deny these obvious points. They seem to suppose that reasoned revision is or ought always to be in accordance with a special principle of "conditionalization" that applies when one comes to treat evidence E as certain. The claim is that in such a case one is to modify one's other degrees of belief so that the new probability one assigns to any given proposition P is given by the following formula:

$$\text{new prob } (P) = \frac{\text{old prob } (P\&E)}{\text{old prob } (E)}$$

The quotient on the right-hand side is sometimes called the conditional probability of P given E, which is why the principle is called conditionalization.

R. C. Jeffrey (1983, chap. 11) shows how this formula can be generalized to allow for the case in which evidence propositions change in probability without becoming certain. Suppose that there are n relevant atomic evidence propositions $E_1, \ldots E_n$, so that there are 2^n strongest conjunctions C_i, each containing E_i, or its denial. Then the new probability one assigns to any given proposition P is the sum of all the quantities of the following form:

$$\text{new prob } (C_i) \times \frac{\text{old prob } (P\&C_i)}{\text{old prob}(C_i)}$$

So, let us consider the following hypothesis, which is widely accepted by subjective probability theorists:

> *Reasoning is conditionalization* The updating of probabilities via conditionalization or generalized conditionalization is (or ought to be) the only principle of reasoned revision.

One way to argue for this is to try to show that various intuitively acceptable principles of reasoning from evidence can be accounted for if this hypothesis is accepted (e.g., Dorling 1972; Horwich 1982).

However, there is a problem with making extensive use of this method of updating. One can use conditionalization to get a new probability for P only if one has already assigned a prior probability not only to E but to P & E. If one is to be prepared for various possible conditionalizations, then for every proposition P one wants to update, one must already have assigned probabilities to various conjunctions of P together with one or more of the possible evidence propositions and/or their denials. Unhappily, this leads to a combinatorial explosion, since the number of such conjunctions is an exponential function of the number of possibly relevant evidence propositions. In other words, to be prepared for

coming to accept or reject any of ten evidence propositions, one would have to record probabilities of over a thousand such conjunctions for each proposition one is interested in updating. To be prepared for twenty evidence propositions, one must record a million probabilities. For thirty evidence propositions, a billion probabilities are needed, and so forth.

Clearly, one could not represent all the needed conjunctions explicitly. One would have to represent them implicitly using some sort of general principle. Given such a general principle, one's total probability distribution would then be determined, by either (1) the total evidence one accepts as certain (using conditionalization) or (2) the various new probabilities assigned to the C_i, (using Jeffrey's generalization of conditionalization). But neither (1) nor (2) is feasible. Consider what is involved in each case.

The idea behind (1) is to represent the degrees of belief to which one is presently committed by means of some general principle, specifying an initial probability distribution, together with a list of all the evidence one has come to treat as certain. Such evidence will include all immediate perceptual evidence – how things look, sound, smell, etc., to one at this or that moment. One will have to remember all such evidence that has influenced one's present degrees of belief. But in fact one rarely remembers such evidence beyond the moment in which one possesses it (a point I return to [in the following part of this chapter]). So (1) is not a usable approach.

On the other hand, (2) requires that one keep track of one's current degree of belief in each of the relevant conjunctions C_i, of evidence propositions and/or their denials. These are things one does not have to be certain about, so the relevant propositions need not be for the most part about immediate perceptual experience, as in (1). So the objection that one hardly ever remembers such propositions does not apply to (2). But (2) is also unworkable, since the number of relevant conjunctions C_i is an exponential function of the number of atomic evidence propositions.

These objections assume one sticks with one's original general principle describing one's initial degrees of belief and records one's present degrees of belief by representing the new evidence accepted as certain or the new probabilities of the various conjunctions C_i.

Alternatively, one might try each time to find a new principle describing one's updated degrees of belief in a single general statement. But the problem of finding such a general principle is intractable, and anyway there will normally be no simpler way to describe one's new probability distribution than the description envisioned in (1) or (2), so this will not normally be feasible either.

Doing extensive updating by conditionalization or generalized conditionalization would be too complicated in practice. Therefore one must follow other principles in revising one's views. It is *conceivable* that all or some of these principles might refer to strength or degree of belief and not just to whether one believes something in a yes/no fashion. But the actual principles we follow do not seem to be of that sort, and it is unclear how these principles might be modified to be sensitive to degree or strength of belief. In the rest of this book I assume that, as far as the principles of revision we follow are concerned, belief is an all-or-nothing matter. I assume that this is so because it is too complicated for mere finite beings to make extensive use of probabilities.

POSITIVE VERSUS NEGATIVE UNDERMINING

I now want to compare two competing theories of reasoned belief revision, which I will call the foundations theory and the coherence theory since they are similar to certain philosophical theories of justification sometimes called foundations and coherence theories (Sosa 1980; Pollock 1979). But the theories I am concerned with are not precisely the same as the corresponding philosophical theories of justification, which are not normally presented as theories of belief revision. Actually, I am not sure what these philosophical theories of "justification" are supposed to be concerned with. So, although I will be using the *term* "justification" in what follows, as well as the terms "coherence" and "foundations," I do not claim that my use of any of these terms is the same as its use in these theories of justification. I mean to be raising a new issue, not discussing an old one.

The key issue is whether one needs to keep track of one's original justifications for beliefs. What I am calling the *foundations* theory says yes; what I am calling the *coherence* theory says no.

The foundations theory holds that some of one's beliefs "depend on" others for their current justification; these other beliefs may depend on still others, until one gets to foundational beliefs

that do not depend on any further beliefs for their justification. In this view reasoning or belief revision should consist, first, in subtracting any of one's beliefs that do not now have a satisfactory justification and, second, in adding new beliefs that either need no justification or are justified on the basis of other justified beliefs one has.

On the other hand, according to the coherence theory, it is not true that one's ongoing beliefs have or ought to have the sort of justificational structure required by the foundations theory. In this view ongoing beliefs do not usually require any justification. Justification is taken to be required only if one has a special reason to doubt a particular belief. Such a reason might consist in a conflicting belief or in the observation that one's beliefs could be made more "coherent," that is, more organized or simpler or less ad hoc, if the given belief were abandoned (and perhaps if certain other changes were made). According to the coherence theory, belief revision should involve minimal changes in one's beliefs in a way that sufficiently increases overall coherence.

In this chapter I elaborate these two theories in order to compare them with actual reasoning and intuitive judgments about such reasoning. It turns out that the theories are most easily distinguished by the conflicting advice they occasionally give concerning whether one should *give up* a belief *P* from which many other of one's beliefs have been inferred, when *P*'s original justification has to be abandoned. Here a surprising contrast seems to emerge – "is" and "ought" seem to come apart. The foundations theory seems, at least at first, to be more in line with our intuitions about how people *ought* to revise their beliefs; the coherence theory is more in line with what people *actually do* in such situations. Intuition seems strongly to support the foundations theory over the coherence theory as an account of what one is *justified* in doing in such cases; but *in fact* one will tend to act as the coherence theory advises.

After I explain this I consider how this apparent discrepancy can be resolved. I conclude that the coherence theory is normatively correct after all, despite initial appearances.

The Foundations Theory of Belief Revision

The basic principle of the foundations theory, as I will interpret it, is that one must keep track of one's original reasons for one's beliefs, so that one's ongoing beliefs have a justificational structure, some beliefs serving as reasons or justifications for others. These justifying beliefs are more basic or fundamental for justification than the beliefs they justify.

The foundations theory rejects any principle of *conservatism*. In this view a proposition cannot acquire justification simply by being believed. The justification of a given belief cannot be, either in whole or in part, that one has that belief. For example, one's justification for believing something cannot be that one already believes it and that one's beliefs in this area are reliable.

Justifications are *prima facie* or defeasible. The foundations theory allows, indeed insists, that one can be justified in believing something *P* and then come to believe something else that undermines one's justification for believing *P*. In that case one should stop believing *P*, unless one has some further justification that is not undermined.

I say "unless one has some further justification," because in this view a belief can have more than one justification. To be justified, a belief must have *at least* one justification. That is, if a belief in *P* is to be justified, it is required either that *P* be a foundational belief whose intrinsic justification is not defeated or that there be at least one undefeated justification of *P* from other beliefs one is justified in believing. If one believes *P* and it happens that all one's justifications for believing *P* come to be defeated, one is no longer justified in continuing to believe *P*, and one should subtract *P* from one's beliefs.

Furthermore, and this is important, if one comes not to be justified in continuing to believe *P* in this way, then not only is it true that one must abandon belief in *P* but justifications one has for other beliefs are also affected if these justifications appeal to one's belief in *P*. Justifications appealing to *P* must be abandoned when *P* is abandoned. If that means further beliefs are left without justification, then these beliefs too must be dropped along with any justifications appealing to them. So there will be a chain reaction when one loses justification for a belief on which other beliefs depend for their justification. (This is worked out in more detail for an artificial intelligence system by Doyle (1979, 1980).)

Now, it is an important aspect of the foundations theory of reasoning that justifications cannot legitimately be circular. *P* cannot be part of the justification for Q while Q is part of the justification for *P* (unless one of these beliefs has a different justification that does not appeal to the other belief).

The foundations theory also disallows infinite justifications. It does not allow *P* to be justified by appeal to Q, which is justified by appeal to *R*, and so on forever. Since justification cannot be circular, justification must eventually end in beliefs that either need no justification or are justified but not by appeal to other beliefs. Let us say that such basic or foundational beliefs are intrinsically justified.

For my purposes it does not matter exactly which beliefs are taken to be intrinsically justified in this sense. Furthermore, I emphasize that the foundations theory allows for situations in which a basic belief has its intrinsic justification defeated by one or more other beliefs, just as it allows for situations in which the justification of one belief in terms of other beliefs is defeated by still other beliefs. As I am interpreting it, foundationalism is not committed to the *incorrigibility* of basic beliefs.

A belief is a basic belief if it has an intrinsic justification which does not appeal to other beliefs. A basic belief can also have one or more nonintrinsic justifications which do appeal to other beliefs. So, a basic belief can have its intrinsic justification defeated and still remain justified as long as it retains at least one justification that is not defeated.

The existence of basic beliefs follows from the restrictions against circular and infinite justifications. Infinite justifications are to be ruled out because a finite creature can have only a finite number of beliefs, or at least only a finite number of *explicit beliefs*, whose content is explicitly represented in the brain. What one is justified in believing either implicitly or explicitly depends entirely on what one is justified in believing explicitly. To consider whether one's implicit beliefs are justified is to consider whether one is justified in believing the explicit beliefs on which the implicit beliefs depend. A justification for a belief that appeals to other beliefs must always appeal to things one believes explicitly. Since one has only finitely many explicit beliefs, there are only finitely many beliefs that can be appealed to for purposes of justification, and so infinite justifications are ruled out.

The Coherence Theory of Belief Revision

The coherence theory is *conservative* in a way the foundations theory is not. The coherence theory supposes one's present beliefs are justified just as they are in the absence of special reasons to change them, where changes are allowed only to the extent that they yield sufficient increases in coherence. This is a striking difference from the foundations theory. The foundations theory says one is justified in continuing to believe something only if one has a special reason to continue to accept that belief, whereas the coherence theory says one is justified in continuing to believe something as long as one has no special reason to stop believing it.

According to the coherence theory, if one's beliefs are incoherent in some way, because of outright inconsistency or simple *ad hoc*ness, then one should try to make minimal changes in those beliefs in order to eliminate the incoherence. More generally, small changes in one's beliefs are justified to the extent these changes add to the coherence of one's beliefs.

For present purposes, I do not need to be too specific as to exactly what coherence involves, except to say it includes not only consistency but also a network of relations among one's beliefs, especially relations of implication and explanation.

It is important that coherence competes with conservatism. It is as if there were two aims or tendencies of reasoned revision, to maximize coherence and to minimize change. Both tendencies are important. Without conservatism a person would be led to reduce his or her beliefs to the single Parmenidean thought that all is one. Without the tendency toward coherence we would have what Peirce (1877) called the method of tenacity, in which one holds to one's initial convictions no matter what evidence may accumulate against them.

According to the coherence theory, the assessment of a challenged belief is always holistic. Whether such a belief is justified depends on how well it fits together with everything else one believes. If one's beliefs are coherent, they are mutually supporting. All one's beliefs are, in a sense, equally fundamental. In the coherence theory there are not the asymmetrical justification relations among one's ongoing beliefs that there are in the foundations theory. It can happen in the coherence theory that *P* is justified because of the way it coheres with Q and Q is justified because of the way it coheres with *P*. In the foundations theory, such a pattern of justification is ruled out by the restriction against circular justification. But there is nothing wrong with circular justification in the coherence theory, especially if the circle is a large one!

I turn now to testing the foundations and coherence theories against our intuitions about cases. This raises an apparent problem for the coherence theory.

An Objection to the Coherence Theory: Karen's Aptitude Test

Sometimes there clearly are asymmetrical justification relations among one's beliefs.

Consider Karen, who has taken an aptitude test and has just been told her results show she has a considerable aptitude for science and music but little aptitude for history and philosophy. This news does not correlate perfectly with her previous grades. She had previously done well not only in physics, for which her aptitude scores are reported to be high, but also in history, for which her aptitude scores are reported to be low. Furthermore, she had previously done poorly not only in philosophy, for which her aptitude scores are reported to be low, but also in music, for which her aptitude scores are reported to be high.

After carefully thinking over these discrepancies, Karen concludes that her reported aptitude scores accurately reflect and are explained by her actual aptitudes; so she has an aptitude for science and music and no aptitude for history and philosophy; therefore her history course must have been an easy one, and also she did not work hard enough in the music course. She decides to take another music course and not to take any more history.

It seems quite clear that, in reaching these conclusions, Karen bases some of her beliefs on others. Her belief that the history course was easy depends for its justification on her belief that she has no aptitude for history, a belief which depends in turn for its justification on her belief that she got a low score in history on her aptitude test. There is no dependence in the other direction. For example, her belief about her aptitude test score in history is not based on her belief that she has no aptitude for history or on her belief that the history course was an easy one.

According to the coherence theory, the relevant relations here are merely *temporal* or *causal* relations. The coherence theory can agree that Karen's belief about the outcome of her aptitude test precedes and is an important cause of her belief that the history course she took was an easy one. But the coherence theory denies that a relation of dependence or justification holds or ought to hold between these two beliefs as time goes by, once the new belief has been firmly accepted.

In order to test this, let me tell more of Karen's story. Some days later she is informed that the report about her aptitude scores was incorrect! The scores reported were those of someone else whose name was confused with hers. Unfortunately, her own scores have now been lost. How should Karen revise her views, given this new information?

The foundations theory says she should abandon all beliefs whose justifications depend in part on her prior belief about her aptitude test scores. The only exception is for beliefs for which she can now find another and independent justification which does not depend on her belief about her aptitude test scores. She should continue to believe only those things she would have been justified in believing if she had never been given the false information about those scores. The foundations theory says this because it does not accept a principle of conservatism. The foundations theory does not allow that a belief can acquire justification simply by being believed.

Let us assume that, if Karen had not been given the false information about her aptitude test scores, she could not have reasonably reached any of the conclusions she did reach about her aptitudes for physics, history, philosophy, and music; and let us also assume that without those beliefs Karen could not have reached any of her further conclusions about the courses she has already taken. Then, according to the foundations theory, Karen should abandon her beliefs about her relative aptitudes for these subjects, and she should give up her belief that the history course she took was easy as well as her belief that she did not work hard enough in the music course. She should also reconsider her decisions to take another course in music and not to take any more history courses.

The coherence theory does not automatically yield the same advice that the foundations theory gives about this case. Karen's new information does produce a loss of overall coherence in her beliefs, since she can no longer coherently suppose that her aptitudes for science, music, philosophy, and history are in any way responsible for the original report she received about the results of her aptitude test. She must abandon that particular supposition about the explanation of the original report of her scores. Still, there is considerable coherence among the beliefs she inferred from this false report. For example, there is a connection between her belief that she has little aptitude for history, her belief that her high grade in the history course was the result of the course's being an easy one, and her belief that she will not take any more courses in history. There are similar connections between her beliefs about her aptitudes

for other subjects, how well she did in courses in those subjects, and her plans for the future in those areas. Let us suppose that from the original report Karen inferred a great many other things that I haven't mentioned; so there are many beliefs involved here. Abandoning all these beliefs is costly from the point of view of conservatism, which says to minimize change. Suppose that there are so many of these beliefs and that they are so connected with each other and with other things Karen believes that the coherence theory implies Karen should retain all these new beliefs even though she must give up her beliefs about the explanation of the report of her aptitude scores. (In fact, we do not really need to suppose these beliefs are intricately connected with each other or even that there are many of them, since in the coherence theory a belief *does* acquire justification simply by being believed.)

The foundations theory says Karen should give up all these beliefs, whereas the coherence theory says Karen should retain them. Which theory is right about what Karen ought to do? Almost everyone who has considered this issue sides with the foundations theory: Karen should not retain any beliefs she inferred from the false report of her aptitude test scores that she would not have been justified in believing in the absence of that false report. That does seem to be the intuitively right answer. The foundations theory is in accordance with our intuitions about what Karen *ought* to do in a case like this. The coherence theory is not.

Belief Perseverance

In fact, Karen would almost certainly keep her new beliefs! That is what people actually do in situations like this. Although the foundations theory seems to give intuitively satisfying advice about what Karen *ought* to do in such a situation, the coherence theory is more in accord with what people actually do.

To document the rather surprising facts here, let me quote at some length from a recent survey article (Ross and Anderson 1982, pp. 147–149), which speaks of

> the dilemma of the social psychologist who has made use of deception in the course of an experiment and then seeks to debrief the subjects who had been the target of such deception. The psychologist reveals the totally contrived and inauthentic nature of the information presented presuming that this debriefing will thereby eliminate any effects such information might have exerted upon the subjects' feelings or beliefs. Many professionals, however, have expressed public concern that such experimental deception may do great harm that is not fully undone by conventional debriefing procedures. . . .

Ross and Anderson go on to describe experiments designed to "explore" what they call "the phenomenon of belief perseverance in the face of evidential discrediting." In one experiment.

> Subjects first received continuous false feedback as they performed a novel discrimination task (i.e., distinguishing authentic suicide notes from fictitious ones). . . . [Then each subject] received a standard debriefing session in which he learned that his putative outcome had been predetermined and that his feedback had been totally unrelated to actual performance. . . . [E]very subject was led to explicitly acknowledge his understanding of the nature and purpose of the experimental deception. Following this total discrediting of the original information, the subjects completed a dependent variable questionnaire dealing with [their] performance and abilities. The evidence for postdebriefing impression perseverance was unmistakable. . . . On virtually every measure . . . the totally discredited initial outcome manipulation produced significant "residual" effects upon [subjects'] . . . assessments. . . .
>
> Follow-up experiments have since shown that a variety of unfounded personal impressions, once induced by experimental procedures, can survive a variety of total discrediting procedures. For example, Jennings, Lepper, and Ross . . . have demonstrated that subjects' impressions of their ability at interpersonal persuasion (having them succeed or fail to convince a confederate to donate blood) can persist after they have learned that the initial outcome was totally inauthentic. Similarly, . . . two related experiments have shown that students' erroneous impressions of their "logical problem solving abilities" (and their academic choices in a follow-up measure two months later) persevered even after they had learned that good or poor teaching procedures provided a totally sufficient explanation for the successes or failures that were the basis for such impressions.
>
> . . . [Other] studies first manipulated and then attempted to undermine subjects'

> theories about the functional relationship between two measured variables: the adequacy of firefighters' professional performances and their prior scores on a paper and pencil test of risk performance. . . . [S]uch theories survived the revelations that the cases in question had been totally fictitious and the different subjects had, in fact, received opposite pairings of riskiness scores and job outcomes. . . . [O]ver 50% of the initial effect of the "case history" information remained after debriefing.
>
> In summary, it is clear that beliefs can survive . . . the total destruction of their original evidential bases.

It is therefore quite likely that Karen will continue to believe many of the things she inferred from the false report of her aptitude test scores. She will continue to believe these things even after learning that the report was false.

The Habit Theory of Belief

Why is it so hard for subjects to be debriefed? Why do people retain conclusions they have drawn from evidence that is now discredited? One possibility is that belief is a kind of habit. This is an implication of behaviorism, the view that beliefs and other mental attitudes are habits of behavior. But the suggestion that beliefs are habits might be correct even apart from behaviorism. The relevant habits need not be overt behavioral habits. They might be habits of thought. Perhaps, to believe that *P* is to be disposed to *think* that *P* under certain conditions, to be disposed to use this thought as a premise or assumption in reasoning and in deciding what to do. Then, once a belief has become established, considerable effort might be needed to get rid of it, even if the believer should come to see that he or she ought to get rid of it, just as it is hard to get rid of other bad habits. One can't simply decide to get rid of a bad habit; one must take active steps to ensure that the habit does not reassert itself. Perhaps it is just as difficult to get rid of a bad belief.

Goldman (1978) mentions a related possibility, observing that Anderson and Bower (1973) treat coming to believe something as the establishing of connections, or "associative links," between relevant conceptual representations in the brain. Now, it may be that, once set up, such connections or links cannot easily be broken unless competing connections are set up that overwhelm the original ones. The easiest case might be that in which one starts by believing *P* and then comes to believe *not P* by setting up stronger connections involving *not P* than those involved in believing *P*. It might be much harder simply to give up one's belief in *P* without substituting a contrary belief. According to this model of belief, in order to stop believing *P*, it would not be enough simply to notice passively that one's evidence for *P* had been discredited. One would have to take positive steps to counteract the associations that constitute one's belief in *P*. The difficulties in giving up a discredited belief would be similar in this view to the difficulties envisioned in the habit theory of belief.

But this explanation does not give a plausible account of the phenomenon of belief perseverance. Of course, there are cases in which one has to struggle in order to abandon a belief one takes to be discredited. One finds oneself coming back to thoughts one realizes one should no longer accept. There are such habits of thought, but this is not what is happening in the debriefing studies. Subjects in these studies are not struggling to abandon beliefs they see are discredited. On the contrary, the subjects do not see that the beliefs they have acquired have been discredited. They come up with all sorts of "rationalizations" (as we say) appealing to connections with other beliefs of a sort that the coherence theory, but not the foundations theory, might approve. So the correct explanation of belief perseverance in these studies is not that beliefs which have lost their evidential grounding are like bad habits.

Positive versus Negative Undermining

In fact, what the debriefing studies show is that people simply do not keep track of the justification relations among their beliefs. They continue to believe things after the evidence for them has been discredited because they do not realize what they are doing. They do not understand that the discredited evidence was the *sole* reason why they believe as they do. They do not see they would not have been justified in forming those beliefs in the absence of the now discredited evidence. They do not realize these beliefs have been undermined. It is this, rather than the difficulty of giving up bad habits, that is responsible for belief perseverance.

The foundations theory says people should keep track of their reasons for believing as they do and should stop believing anything that is not associated with adequate evidence. So the foundations theory implies that, if Karen has not kept track of her reason for believing her history

course was an easy one, she should have abandoned her belief even before she was told about the mix-up with her aptitude test scores. This seems clearly wrong.

Furthermore, since people rarely keep track of their reasons, the theory implies that people are unjustified in almost all their beliefs. This is an absurd result! The foundations theory turns out not to be a plausible normative theory after all. So let us see whether we cannot defend the coherence theory as a normative theory.

We have already seen how the coherence theory can appeal to a nonholistic *causal* notion of local justification by means of a limited number of one's prior beliefs, namely, those prior beliefs that are most crucial to one's justification for adding the new belief. The coherence theory does not suppose there are *continuing* links of justification dependency that can be consulted when revising one's beliefs. But the theory can admit that Karen's coming to believe certain things depended on certain of her prior beliefs in a way that it did not depend on others, where this dependence represents a kind of local justification, even though in another respect whether Karen was justified in coming to believe those things depended on everything she then believed.

Given this point, I suggest the coherence theory can suppose it is incoherent to believe both *P* and also that all one's reasons for believing *P* relied crucially on false assumptions. Within the coherence theory, this implies, roughly, the following:

> *Principle of Positive Undermining* One should stop believing *P* whenever one positively believes one's reasons for believing *P* are no good.

This is only roughly right, since there is also the possibility that one should instead stop believing that one's reasons for *P* are no good, as well as the possibility that one cannot decide between that belief and *P*. In any event, I want to compare this rough statement of the principle with the corresponding principle in a foundations theory:

> *Principle of Negative Undermining* One should stop believing *P* whenever one does not associate one's belief in *P* with an adequate justification (either intrinsic or extrinsic).

The Principle of Positive Undermining is much more plausible than the Principle of Negative Undermining. The Principle of Negative Undermining implies that, as one loses track of the justifications of one's beliefs, one should give up those beliefs. But, if one does not keep track of one's justifications for most of one's beliefs, as seems to be the case, then the Principle of Negative Undermining says that one should stop believing almost everything one believes, which is absurd. On the other hand the Principle of Positive Undermining does not have this absurd implication. The Principle of Positive Undermining does not suppose that the absence of a justification is a reason to stop believing something. It only supposes that one's belief in *P* is undermined by the *positive* belief that one's reasons for *P* are no good.

It is relevant that subjects *can* be successfully debriefed after experiments involving deception if they are made vividly aware of the phenomenon of belief perseverance, that is, if they are made vividly aware of the tendency for people to retain false beliefs after the evidence for them has been undercut, and if they are also made vividly aware of how this phenomenon has acted in their own case (Nisbett and Ross 1980, p. 177). It might be suggested that this shows that under ideal conditions people really do act in accordance with the foundations theory after all, so that the foundations theory *is* normatively correct as an account of how one ideally ought to revise one's beliefs. But in fact this further phenomenon seems clearly to support the coherence theory, with its Principle of Positive Undermining, and not the foundations theory, with its Principle of Negative Undermining. The so-called process debriefing cannot merely undermine the evidence for the conclusions subjects have reached but must also directly attack each of these conclusions themselves. Process debriefing works not just by getting subjects to give up beliefs that originally served as evidence for the conclusions they have reached but by getting them to accept certain further positive beliefs about their lack of good reasons for each of these conclusions.

What about Our Intuitions?

It may seem to fly in the face of common sense to suppose that the coherence theory is normatively correct in cases like this. Remember that, after carefully considering Karen's situation, almost everyone agrees she should give up all beliefs inferred from the original false report, except those beliefs which would have been justified apart from any appeal to evidence tainted by that false information. Almost everyone's judgment about what Karen ought to do coincides with what the foundations theory

says she ought to do. Indeed, psychologists who have studied the phenomenon of belief perseverance in the face of debriefing consider it to be a paradigm of irrationality. How can these strong normative intuitions possibly be taken to be mistaken, as they must be if the coherence theory is to be accepted as normatively correct?

The answer is that, when people think about Karen's situation, they ignore the possibility that she may have failed to keep track of the justifications of her beliefs. They imagine Karen is or ought to be aware that she no longer has any good reasons for the beliefs she inferred from the false report. And, of course, this is to imagine that Karen is violating the Principle of Positive Undermining. It is hard to allow for the possibility that she may be violating not that principle but only the foundationalist's Principle of Negative Undermining.

Keeping Track of Justification

People do not seem to keep track of the justifications of their beliefs. If we try to suppose that people do keep track of their justifications, we would have to suppose that either they fail to notice when their justifications are undermined or they do notice but have great difficulty in abandoning the unjustified beliefs in the way a person has difficulty abandoning a bad habit. Neither possibility offers a plausible account of the phenomenon of belief perseverance.

It stretches credulity to suppose people always keep track of the sources of their beliefs but often fail to notice when these sources are undermined. That is like supposing people always remember everything that has ever happened to them but cannot always retrieve the stored information from memory. To say one remembers something is to say one has stored it in a way that normally allows it to be retrieved at will. Similarly, to say people keep track of the sources of their beliefs must be to say they can normally use this information when it is appropriate to do so.

I have already remarked that the other possibility seems equally incredible, namely, that people have trouble abandoning the undermined beliefs in the way they have trouble getting rid of bad habits. To repeat, participants in belief perseverance studies show no signs of knowing their beliefs are ungrounded. They do not act like people struggling with their beliefs as with bad habits. Again, I agree it sometimes happens that one keeps returning to thoughts after one has seen there can be no reason to accept those thoughts. There are habits of thought that can be hard to get rid of. But that is not what is going on in the cases psychologists study under the name of belief perseverance.

This leaves the issue of whether one should *try* always to keep track of the local justifications of one's beliefs, even if, in fact, people do not seem to do this. I want to consider the possibility that there is a good reason for not keeping track of these justifications.

Clutter Avoidance Again

We have seen there is a practical reason to avoid too much clutter in one's beliefs. There is a limit to what one can remember, a limit to the number of things one can put into long-term storage, and a limit to what one can retrieve. It is important to save room for important things and not clutter one's mind with a lot of unimportant matters. This is an important reason why one does not try to believe all sorts of logical consequences of one's beliefs. One should not try to infer all one can from one's beliefs. One should try not to retain too much trivial information. Furthermore, one should try to store in long-term memory only the key matters that one will later need to recall. When one reaches a significant conclusion from one's other beliefs, one needs to remember the conclusion but does not normally need to remember all the intermediate steps involved in reaching that conclusion. Indeed, one should not try to remember those intermediate steps; one should try to avoid too much clutter in one's mind.

Similarly, even if much of one's knowledge of the world is inferred ultimately from what one believes oneself to be immediately perceiving at one or another time, one does not normally need to remember these original perceptual beliefs or many of the various intermediate conclusions drawn from them. It is enough to recall the more important of one's conclusions. This means one should not be disposed to try to keep track of the local justifications of one's beliefs. One could keep track of these justifications only by remembering an incredible number of mostly perceptual original premises, along with many, many intermediate steps which one does not want and has little need to remember. One will not want to link one's beliefs to such justifications because one will not in general want to try to retain the prior beliefs from which one reached one's current beliefs.

The practical reason for not keeping track of the justifications of one's beliefs is not as severe

as the reason that prevents one from trying to operate purely probabilistically, using generalized conditionalization as one's only principle of reasoned revision. The problem is not that there would be a combinatorial explosion. Still, there are important practical constraints. It is more efficient not to try to retain these justifications and the accompanying justifying beliefs. This leaves more room in memory for important matters.

Bibliography

Anderson, J. R., and Bower, G. H. (1973). *Human Associative Memory* (Washington, D.C.: Winston).

Dorling, Jon (1972). "Bayesianism and the rationality of scientific inference," *British Journal for the Philosophy of Science* 23:181–190.

Doyle, Jon (1979). "A truth maintenance system," *Artificial Intelligence* 12:231–272.

Doyle, Jon (1980). "A Model for Deliberation, Action, and Introspection," MIT Artificial Intelligence Laboratory Technical Report 581.

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

Goldman, Alvin I. (1978). "Epistemology and the psychology of belief," *Monist* 61:525–535.

Horwich, Paul (1982). *Probability and Evidence* (Cambridge: Cambridge University Press).

Jeffrey, Richard C. (1983). *The Logic of Decision* (Chicago: University of Chicago Press).

Kyburg, Henry (1961). *Probability and the Logic of Rational Belief* (Middletown, Conn.: Wesleyan University Press).

Nisbett, Richard, and Ross, Lee (1980). *Human Inference: Strategies and Shortcomings of Social Judgement* (Englewood Cliffs, N.J.: Prentice-Hall).

Peirce, C. S. (1877). "The fixation of belief," *Popular Science Monthly* 12:1–15. Reprinted in *Philosophical Writings of Peirce*, Justice Buchler, ed. (New York: Dover, 1955), 5–22.

Pollock, John (1979). "A plethora of epistemological theories," in *Justification and Knowledge*, George Pappas, ed. (Dordrecht, Holland: Reidel), 93–114.

Ross, Lee, and Anderson, Craig A. (1982). "Shortcomings in the attribution process: On the origins and maintenance of erroneous social assessments," in *Judgement under Certainty: Heuristics and Biases*, Daniel Kahneman, Paul Slovic, and Amos Tversky, eds. (Cambridge: Cambridge University Press), 129–152.

Sosa, Ernest (1980). "The raft and the pyramid: Coherence versus foundations in the theory of knowledge," *Midwest Studies in Philosophy* 5:3–25.