***Articulating Reasons*: Chapter Three**

Insights and Blindspots of Reliabilism

I

One of the most important developments in the theory of knowledge during the past two decades has been a shift in emphasis to concern with issues of the *reliability* of various processes of belief formation. One way of arriving at beliefs is more reliable than another in a specified set of circumstances just insofar as it is more *likely*, in those circumstances, to produce a *true* belief. Classical epistemology, taking its cue from Plato, understood knowledge as justified true belief. While Gettier had raised questions about the joint *sufficiency* of those three conditions, it is only more recently that their individual *necessity* was seriously questioned. What I will call the 'Founding Insight' of reliabilist epistemologies is the claim that true beliefs can, at least in some cases, amount to genuine knowledge even where the justification condition is not met (in the sense that the candidate knower is unable to produce suitable justifications), provided the beliefs resulted from the exercise of capacities that are *reliable* producers of true beliefs in the circumstances in which they were in fact exercised.

The original motivation for the justification leg of the JTB epistemological tripod—for, in Plato's terminology, taking knowledge to require true opinion *plus an account*—is that merely *accidentally* true beliefs do not generally qualify as cases of knowledge. The man who guesses correctly which road leads to Athens, or who acquires his belief by flipping a coin, should not be said to *know* which is the correct road, even in the cases where he happens to be right. A space is cleared for reliabilism by the observation that supplying *evidence* for a claim, offering *reasons* for it, *justifying* it, are

not the only ways in which to show that a belief is, if true, not true merely by accident. For that it suffices to show that the belief is of a kind that could, under the prevailing circumstances, have been *expected* or *predicted* to be true.[1] That the believer possesses good reasons for the belief is only one basis for such an expectation or prediction.

Consider an expert on classical Central American pottery who over the years has acquired the ability to tell Toltec from Aztec potsherds—reliably though not infallibly—simply by looking at them. We may suppose that there are no separately distinguishing features of the fragments that she can cite in justifying her classifications. When looking closely at the pieces, she just finds herself believing that some of them are Toltec, and others Aztec. Suppose further that she regards beliefs formed in this way with great suspicion; she is not willing to put much weight on them, and in particular is not willing to risk her professional reputation on convictions with this sort of provenance. Before reporting to colleagues, or publishing conclusions that rest on evidence as to whether particular bits are Toltec or Aztec, she always does microscopic and chemical analyses that give her solid inferential evidence for the classification. That is, she does not believe that she is a reliable noninferential reporter of Toltec and Aztec potsherds; she insists on confirmatory evidence for beliefs on this topic that she has acquired noninferentially. But suppose that her colleagues, having followed her work over the years, have noticed that she is in fact a reliable distinguisher of one sort of pottery from the other. Her off-the-cuff inclinations to call something Toltec rather than Aztec can be trusted. It seems reasonable for them to say, in some case where she turned out to be right, that although she insisted on confirmatory evidence for her belief, in fact she *already knew* that the

[1] The expectation or prediction need not rise to the level of perfect *certainty*. Although there may well be a use of 'knows' that requires such certainty, it was one of the great advances in twentieth century epistemology prior to the advent of reliabilism to realize that such a concept of knowledge not only includes an unrefusable invitation to skepticism, it is of no use for discussing the achievements of science, and in any case is not obligatory. If our ordinary use of 'know' involves such commitments, that is the best possible reason to replace it by a less committive technical notion that is more useful for our central epistemological purposes. The fact that there are circumstances in which we would have been wrong should not preclude our counting as knowing in the cases where we are in fact right. Our fallibility should not be taken to rule out the possibility of knowledge.

fragment in question was Toltec, even before bringing her microscope and reagents into play.[2]

If that is the right thing to say about a case of this sort, then knowledge attributions can be underwritten by a believer's *reliability*, even when the believer is not in a position to offer *reasons* for the belief. If they can be so underwritten, then justificatory internalism in epistemology is wrong to restrict attributions of knowledge to cases where the candidate knower can offer reasons inferentially justifying her (true) beliefs.[3] Reliabilism is a kind of epistemological *externalism*. For it maintains that facts of which a believer is not aware, and so cannot cite as reasons—e.g. the reliability of her off-the-cuff dispositions to classify potsherds—can make the difference between what she has counting as genuine knowledge and its counting merely as true belief.

So accepting the Founding Insight of reliabilism does involve disagreeing with the verdicts of justificatory internalism in some particular cases. But concern with reliability does not simply *contradict* the genuine insights of classical JTB epistemology. Rather, it can be seen as a *generalization* of the classical account. Reasoning takes its place as one potentially reliable process among others. Accepting only beliefs one *could* give reasons for—even if one did not acquire the belief inferentially by considering such reasons—is, under many circumstances, a reliable technique of belief formation. Where it is *not*, where the two criteria collide, it is arguable that the reliability criterion ought to trump the justificatory one. This might happen where inductive reasons could indeed be given for a belief, but where they are such weak reasons that the inference they

---

[2] According to this line of thought, one can know something without knowing that one knows. (The Kp→KKp principle fails.) One might believe that p without believing that one knows that p. For, as in the example just offered, one may not even believe that the belief is the outcome of a reliable process, though it is. The attitude of the believer might be that the belief she finds herself with perceptually just happened in this case to be true. Since belief is a condition for knowledge, if one does not even *believe* that one knows that p, then one does not *know* that one knows it.

[3] Classical justificatory internalism about knowledge should be taken to require only that the candidate knower could give reasons for her belief, not that the belief in fact have been acquired as the result of antecedent consideration of those reasons. For the stronger requirement would limit knowledge to beliefs acquired inferentially. But we ought to be able to allow that *non*inferentially acquired beliefs—for instance, those acquired perceptually (and, arguably, by memory or even testimony) can constitute genuine knowledge. The requirement would just be that after the fact the believer can offer reasons for her belief—for instance by invoking her own reliability as a noninferential reporter.

underwrite falls short of reliability. Thus a colorful sunset may give some reason to believe the next day will be fine ("Red at night, sailor's delight…"), even though acquiring one's weather beliefs on that basis may be quite unreliable. In such a case, even though one had a reason for what turned out to be a true belief, we might hesitate to say that one *knew* it would not rain. The reliability formula characterizes the role of such sources of knowledge as perception, memory, and testimony—none of which are immediately or obviously inferential in nature—at least as well as, and perhaps better than a characterization of them in terms of looks, memories, and testimony offering *reasons*. That is because those sources *do* provide reasons sufficient for knowledge *at most* in the cases and the circumstances where they are reliable. Unreliable perception, memory, and testimony are *not* sufficient grounds for knowledge (and not for Gettier reasons).

What conclusions about the relations between reliability and reasons follow from what I have called the Founding Insight of reliabilism? The temptation is to suppose that for the reasons just considered, the concept of reliability of belief-forming processes can simply *replace* the concept of having good reasons for belief—that *all* the explanatory work for which we have been accustomed to call on the latter can be performed as well or better by the former. Thinking of things this way is thinking of the Founding Insight as motivating a recentering of epistemology. Classical JTB theories of knowledge had taken as central and paradigmatic exemplars true beliefs that the knower could justify inferentially. Beliefs that were the outcome of reliable processes of belief-formation—for instance the noninferentially arrived at deliverances of sense perception—qualified as special cases of knowledge, just if the believer *knew* (or at least believed) she was a reliable perceiver under those circumstances, and so could cite her reliability as a *reason* for belief. Reliability appeared as just one sort of reason among others. Reliabilist theories of knowledge take as their central and paradigmatic exemplars true beliefs that result from reliable belief-forming mechanisms or strategies, regardless of the capacity of the believer to justify the belief, for instance by citing her reliability. Believing what one

can give reasons for appears as just one sort of reliable belief-forming mechanism among others.

More general theoretical considerations also seem to favor the replacement of the concept of reasons with that of reliability in epistemology. For we ought to ask why the concept of knowledge is of philosophical interest at all. It seems clear why we ought to care about the *truth* of beliefs, both our own and those of others. For the success of our actions often turns on the truth of the beliefs on which they are based.[4] But why should we in addition care about whatever feature distinguishes *knowledge* from mere true belief? Surely it is because we want to be able to *rely* on what others say, to provide us information. This interest in interpersonal communication of information motivates caring about the reliability of the processes that yield a belief, independently of caring about its truth—for we can know something about the one in particular cases without yet knowing about the other. It is not wise to rely on lucky guesses. So independently of the vagaries of the prior epistemological tradition, and independently of how words like 'know' happen to be used in natural languages, we have a philosophical interest in investigating the status of beliefs that are produced by reliable processes. The capacity of a believer to provide reasons for her beliefs seems relevant to this story only at one remove: insofar as it contributes to reliability.

There are three distinguishable questions here. First, do the examples pointed to by the Founding Insight as genuine cases of knowledge stand up to critical scrutiny? For instance, ought we to count our pottery expert as having knowledge in advance of having reasons and in spite of her disbelief in her own reliability? Second, do such examples warrant a recentering of epistemology, to focus on reliability of belief-forming processes rather than possession of reasons as distinctive of the most cognitively significant subclass of true beliefs? Third, does the possibility and advisability of such a recentering

[4] It is tempting to overgeneralize from this platitude (in much the same way as it is from the Founding Insight of reliabilism), by seeking to *define* first the *truth* (and then in turn the *truth conditions*) of beliefs in terms of conduciveness to success of actions based on those beliefs. There are insuperable objections to such an explanatory strategy, which I have discussed in "Unsuccessful Semantics" *Analysis* Vol. 54 No. 3 (July 1994) pp. 175–8.

of epistemology mean that the explanatory role played by the concepts of reasons, evidence, inference, and justification can be taken over by that of reliable belief-forming processes—that is, that they matter *only* as marks of reliability of the beliefs they warrant? The temptation referred to above is the temptation to move from an affirmative answer to the first question to an affirmative answer to the other two. This is a temptation that should be resisted. I am prepared to accept the Founding Insight of reliabilism. But I will present reasons to dispute the recentering of epistemology from reasons to reliability to which it tempts us. And I will present further arguments to reject the replacement of the concept of reasons with that of reliability.

II

To begin with, it is important to realize how delicate and special are the cases to which the Founding Insight appeals. If the expert not only *is* reliable, but *believes* herself to be reliable, then she *does* have a reason for her belief, and issue is not joined with the justificatory internalist. Although the belief was acquired by noninferential perceptual mechanisms, it *could* in that case be justified inferentially. For that the shard is (probably) Toltec follows from the claim that the expert is perceptually disposed to call it 'Toltec', together with the claim that she is reliable in these matters under these circumstances. After all, to take the expert to be reliable just is to take it that the inference from her being disposed to call something 'Toltec' to its being Toltec is a good one. Thus to get a case of knowledge based on reliability without reasons, we need one where a reliable believer does not take or believe herself to be reliable. These are going to be odd cases, since to qualify as even a candidate knower, the individual in question must nonetheless form a belief.

It is not hard to describe situations in which someone in fact reliably responds differentially to some sort of stimulus, without having any idea of the mechanism that is in play. Industrial chicken-sexers can, I am told, reliably sort hatchlings into males and females by inspecting them, without having the least idea how they do it. With enough

training, they just catch on. In fact, as I hear the story, it has been established that although these experts uniformly believe that they make the discrimination visually, research has shown that the cues their discriminations actually depend upon are olfactory. At least in this way of telling the story, they are reliable noninferential reporters of male and female chicks, even though they know nothing about how they can do it, and so are quite unable to offer *reasons* (concerning how it looks or, *a fortiori*, smells) for believing a particular chick to be male. Again, individuals with blindsight are in the ordinary sense blind, and believe that they cannot respond differentially to visual stimuli. Yet they can, at least in some circumstances, reasonably reliably discriminate shapes and colors if forced to guess. Ordinary blindsight phenomena do not yield *knowledge*, since the individuals in question do not come to *believe* that, for instance, there is a red square in front of them. The most they will do is to *say* that, as a guess. For an example relevant to reliabilist concerns, we need a sort of super blindsight. Such super blindsight would be a phenomenon, first, in which the subject is more reliable than is typical for ordinary blindsight. For in the ordinary cases, the most one gets is a statistically significant preponderance of correct guesses relative to chance expectancy. Second, it would be a phenomenon in which the blindsighted individual formed an unaccountable *conviction* or belief that, for instance, there was a red square in front of him. Then we might indeed be tempted, as the Founding Insight urges, to say that the blindsighted individual actually *knew* there was a red square in front of him—just as the naïve chicken-sexer *knows* that he is inspecting a male chick.

But as we saw already in connection with the archeological expert, as so far described, these are cases that can cheerfully be accommodated within the framework of justificatory internalism. For though the examples have been carefully constructed so as to involve mechanisms of belief-acquisition that are themselves *non*inferential, this by itself does not entail that the candidate knowers cannot offer inferential justifications for those beliefs. An epistemological internalism that denied the intelligibility of counting noninferentially acquired beliefs (paradigmatically, those acquired perceptually) as knowledge would be a nonstarter. Perceptual knowledge, according to any JTB account

with any initial plausibility at all, depends on the capacity of the perceiver to offer justifying evidence from which the belief *could* have been inferred, even though in fact that is not how it came about. And the idea of reliability of a belief-forming process is exactly what is required to produce a recipe for such *ex post facto* justifications of noninferentially acquired beliefs.[5] In the standard case, we would expect that a reliable chicken-sexer would come to believe that he is reliable. And that belief, together with his inclination to classify a particular chick as male, provides an appropriate inferential *justification* for the corresponding noninferentially acquired belief. So to put pressure on classical justificatory internalism, we need to build into the case the constraint that the candidate knower, though in fact reliable, does not believe himself to be reliable. This is perhaps most intuitive in the case of blindsight—even super blindsight—since it characteristic of the original phenomenon that the blindsighted continue to insist that they can't see anything. They are, after all, blind.

At this point a tension comes to light. If the expert really does *not* take herself to be a reliable noninferential reporter of Toltec potsherds, one might think that it is cognitively irresponsible of her so much as to form the belief that a particular fragment is Toltec, in advance of her investigation of microscopic and chemical evidence she *does* take to offer reliable indications. If the chicken-sexer does not believe he is a reliable discriminator of male from female chicks (perhaps because he is still early in his training, and does not yet realize that he has caught on), what business does he have coming noninferentially to *believe* that a particular chick is male, as opposed merely to finding himself inclined to say so, or putting it in the bin marked 'M'? Again, *endorsing* that inclination by coming to *believe* the chick is male seems irresponsible at this stage. If the super blindsighted person insists that he is not a reliable reporter of red squares, because he is blind and so cannot *see* red squares, how can he at the same time nonetheless genuinely *believe* that there is a red square in front of him? When thus fully described, are the cases that motivate the Founding Insight still coherent and intelligible?

---

[5] This is Sellars' strategy for defending justificatory internalism, in "Empiricism and the Philosophy of Mind." See the discussion of this point in my Study Guide in *Empiricism and the Philosophy of Mind* by Wilfrid Sellars (Harvard University Press, 1997).

I think that they are. There *is* a certain sort of cognitive irresponsibility involved in those who do not take themselves to be reliable reporters of a certain sort of phenomenon nonetheless coming to believe the reports they find themselves inclined to make. But I do not think that is a decisive reason to deny that it is intelligible to acquire beliefs in this way. Cognitively irresponsible beliefs can genuinely be beliefs. And in these very special cases, such irresponsible beliefs can qualify as knowledge. At the very least, I do not think it is open to the convinced justificatory internalist epistemologist to insist on the incoherence of examples meeting the stringent conditions that have just been rehearsed. For to be "cognitively responsible" in the sense invoked in pointing to the tensions above just means not forming beliefs for which one cannot offer any kind of a reason. Treating examples of the sort sketched above as incoherent is in effect building this requirement into the definition of 'belief'—so that what one has acquired cannot count as a belief unless one is in a position to offer at least some kind of reason for it. To impose that sort of requirement is surely to beg the question against the reliabilist epistemologist.

In fact, there is nothing unintelligible about having beliefs for which we cannot give reasons. Faith—understood broadly as undertaking commitments without claiming corresponding entitlements—is surely not an incoherent concept. (Nor is it by any means the exclusive province of religion.) And should the convictions of the faithful turn out not only to be true, but also (unbeknownst to them) to result from reliable belief-forming processes, I do not see why they should not be taken to constitute knowledge. The proper lesson to draw from the tension involved in the sorts of examples of knowledge to which the Founding Insight draws our attention, I think, is not that those examples are incoherent, but that they are in principle exceptional. Knowledge based on reliability without the subject's having reasons for it[6] is possible as a local phenomenon, but not as a global one.

[6] "Having reasons" rather than "being able to give reasons" because justificatory internalism need not be committed to withholding attributions of knowledge in cases where the possessor of reasons (in the form of other justified beliefs from which the belief in question follows) is *de facto* unable to produce them, say through having forgotten them.

III

For what would it be like for *all* our knowledge, indeed, all our *belief* to be like the examples we have been considering? Granted that cognitively irresponsible belief is possible in special, isolated cases, can we coherently describe practices in which people genuinely have beliefs, but *all* of them are cognitively irresponsible in that they are knowingly held in the absence of reasons for them? Put differently, do belief-forming practices of the sort that motivate the Founding Insight form an autonomous set—that is, a set of practices of belief-formation that one could have though one had no others?

This is an important question in the context of the temptation to understand the significance of the Founding Insight of reliabilism as warranting a *recentering* of epistemology to focus on the reliability of belief-forming processes rather than on possession of reasons as what distinguishes the most philosophically interesting subclass of true beliefs. For the reason-giving practices that the classical justificatory internalist takes as paradigmatic ingredients of knowledge *are* autonomous in this sense. That is, we *can* make sense of a community whose members only formed beliefs when they thought they had justifications for them. Clearly all of their inferentially arrived at beliefs can meet this condition. For the noninferentially acquired beliefs, we must insist only that they form *beliefs* noninferentially only in cases where they believe themselves to be reliable. *Those* beliefs can in turn have been acquired from others (who are and are taken already to be reliable), who train the novices in their discriminations. Thus the children learn reliably to sort lollipops into piles labeled with color words first, and only once certified as reliable noninferential discriminators of colors do they graduate to forming *beliefs* of the form "That lollipop is purple." At that stage, if asked what reasons they have for those beliefs, they can invoke their own reliability. This invocation may be

implicit, consisting for instance in saying something like "I can tell what things are purple by looking at them." They might even say "It looks purple to me," where this need be no more than a code for "I find myself inclined to sort it into the pile labeled 'purple'."[7]

It is at the very least unclear that we can make sense of a community of believers who, while often holding true beliefs, and generally acquiring them by reliable mechanisms, *never* are in a position to offer reasons for their beliefs. This would require that they never take themselves or each other to be reliable. For any attribution of reliability (when conjoined with a claim about what the reliable one believes or is inclined to say) *inferentially* underwrites a conclusion. A community precluded from giving reasons for beliefs cannot so much as have the concept of reliability—nor, accordingly, (by anyone's lights) of knowledge. Its members can serve as measuring instruments—that is, reliable indicators—both of perceptible environing states, and of each other's responses. But they cannot treat themselves or each other as doing that. For they do not discriminate between reliable indication and unreliable indication. Absent such discrimination, they cannot be taken to understand themselves or each other as *indicators* at all. For the very notion of a *correlation* between the states of an instrument and the states that it is a candidate for measuring is unintelligible apart from the assessments of reliability. Though they are reliable indicators, they do not in fact rely on their own or each other's indications, since they draw no conclusions from them.

I think these are good reasons to deny that what such reliable indicators have is knowledge. But the reasons forwarded thus far are at best probative, not dispositive. So far, however, our attention has been focused on the third condition on knowledge: whatever distinguishes it from mere true belief. If we shift our attention to the first condition—the condition that one does not know what one does not *believe*—stronger reasons to doubt the intelligibility of the reliability-without-reasons scenario emerge. For states that do not stand in inferential relations to one another, that do not serve as reasons

[7] For a more nuanced discussion, see my treatment of Sellars account of the logic of 'looks', in the Study Guide in *Empiricism and the Philosophy of Mind*, op. cit..

one for another, are not recognizable as beliefs at all. The world is full of reliable indicators. Chunks of iron rust in wet environments and not in dry ones. Land mines explode when impressed by anything weighing more than a certain amount. Bulls charge red flapping bits of material. And so on. Their reliable dispositions to respond differentially to stimuli, and thereby to sort the stimuli into kinds, do not qualify as *cognitive*, because the responses that are reliably differentially elicited are not applications of *concepts*. They are not the formation of *beliefs*. Why not? What else is required for the reliable responses to count as beliefs? What difference makes the difference between a parrot trained to utter "That's red," in the presence of red things and a genuine noninferential reporter of red things who responds to their visible presence by acquiring the perceptual *belief that* there is something red in front of her?

At a minimum, I want to say, it is the *inferential articulation* of the response. Beliefs—indeed, anything that is propositionally contentful (whose content is in principle specifiable by using a declarative sentence or a 'that' clause formed from one), and so conceptually articulated—are essentially things that can serve as premises and conclusions of inferences. The subject of genuine perceptual beliefs is, as the parrot is not, responding to the visible presence of red things by making a potential move in a game of giving and asking for reasons: applying a concept. The believer is adopting a stance that involves further consequential commitments (for instance, to the object perceived being colored), that is incompatible with other commitments (for instance, to the object perceived being green), and that one can show one's entitlements to in terms of other commitments (for instance, to the object perceived being scarlet). No response that is not a node in a network of such broadly inferential involvements, I claim, is recognizable as the application of *concepts*. And if not, it is not recognizable as a belief, or the expression of a belief, either.

We ought to respect the distinction between genuine perceptual *beliefs*—which require the application of *concepts*—and the reliable responses of minerals, mines, and matador-fodder. I claim that an essential element of that distinction is the potential role

as both premise and conclusion in reasoning (both theoretical and practical) that beliefs play. One might choose to draw this line differently, though I am not aware of a plausible competitor. But I do not think it is open to the reliabilist epistemologist to refuse to draw a line at all. To do that—not merely to broaden somewhat the third condition on knowledge, but to reject the first out of hand—is to change the subject radically. It is not to disagree about the analysis of knowledge, but to insist on talking about something else entirely.[8]

If there is anything to this line of thought, then it is simply a mistake to think that the notion of being reliable could take over the explanatory role played by the notion of having reasons. For what distinguishes propositionally contentful, and therefore conceptually articulated *beliefs*, including those that qualify as knowledge, from the merely reliable responses or representations of noncognitive creatures—those that have know-*how*, but are not in the knowing-*that* line of work—is (at least) that they can both serve as and stand in need of *reasons*. I will call the failure to realize this limitation on the explanatory powers of the concept of reliability the "Conceptual Blindspot" of reliabilism.

That it is a mistake is at base a *semantic* point. But because of the *belief* condition on knowledge it serves to also to temper the conclusions we are entitled to draw from the Founding Insight of *epistemological* reliabilism concerning the *justification* condition. The examples of knowledge based on reliability without the possibility of offering reasons, which motivate the Founding Insight, are *essentially* fringe phenomena. Their intelligibility is parasitic on that of the reason-giving practices that underwrite ordinary ascriptions of knowledge—and indeed, of belief *tout court*. Practices in which some beliefs are accorded the status of true and justified are autonomous—intelligible as games one could play though one played no other. Practices in which the only status beliefs can have besides being true is having been reliably produced are not autonomous in that sense. We must carefully resist the temptation to overstate the significance of the

[8] Semantic programs such as those of Dretske, Fodor, and Millikan are at their weakest when addressing the question of what distinguishes representations that deserve to be called 'beliefs' from other sorts of indicating states.

Founding Insight of reliabilism. Besides serving as a kind of reason, reliability can take a subordinate place alongside reasons in certifying beliefs as knowledge. But it cannot displace giving and asking for reasons from its central place in the understanding of cognitive practice.

IV

So the proper domain of reliabilism is epistemology rather than semantics. Within epistemology, its proper lessons pertain to the condition that distinguishes knowledge from mere true belief. It does not provide the resources to distinguish the genus of which knowledge is a species—*conceptually* articulated, in particular *propositionally* contentful attitudes of *belief*—from the sorts of reliable indication exhibited by reliably indicating artifacts such as measuring instruments. Now perhaps in pointing out that it would be a mistake to treat appeals to reliability as a candidate for replacing appeals to reasons in these broader explanatory domains I am attacking a straw man. The temptations to generalize the lessons of the Founding Insight to which I have been urging resistance may not be widely felt. Insofar as they are not, it would be tendentious to describe the merely notional possibility of such misguided overgeneralizations as constituting a flaw or blindspot in reliabilism itself—once the boundaries of that doctrine are suitable circumscribed.

However it may be with this temptation, there is another that is surely part and parcel of reliabilism's contemporary appeal in epistemology. That is the idea that reliabilism provides at least the raw materials for a *naturalized* epistemology—one that will let us exhibit states of knowledge as products of natural processes fully intelligible in broadly physicalistic terms. The strictures just rehearsed counsel us to care in stating this ambition. Epistemological reliabilism suggests a path whereby *if* and *insofar as* the concept of (propositionally contentful) *belief* can be naturalized, so can the concept of *knowledge*. Reliabilism promises a recipe for extending the one sort of account to the

other. The qualification codified in the antecedent of this conditional is not trivial, but neither is the conditional. In particular, it expresses a claim that convinced justificatory internalists might well have felt obliged to doubt. For if and insofar as what distinguishes knowledge from other true beliefs must be understood in terms of possession of *good reasons* or of justificatory *entitlement* or *warrant*, pessimism about the prospects for eventual naturalistic domestication of these latter normative notions would extend to the concept of knowledge itself.

A belief-forming mechanism is *reliable* (in specified circumstances), just in case it is objectively *likely* (in those circumstances) to result in *true* beliefs. If the notions of *belief* and *truth* have been explained physicalistically or naturalistically[9]—a substantive task, to be sure, but perhaps not a distinctively *epistemological* one—then one of the promises of reliabilism in epistemology is that all one needs to extend those accounts to encompass also *knowledge* is a naturalistic story about objective likelihood. But since it is *objective* likelihood that is at issue—and not subjective matters of conviction or evidence, of what else the subject knows or believes—such a story should not, it seems, be far to seek. For objective probabilities are a staple of explanations in the natural sciences, indeed, even in fundamental physics. The second conclusion the Founding Insight of reliabilism tempts us to draw is accordingly that it provides a recipe for a purely naturalistic account at least of what distinguishes knowledge from other true beliefs.

This line of thought is widely endorsed, even by those who do not applaud the project that motivates it. For it seems to me that even those who reject the premises that form its antecedent accept the conditional that ***if*** the concept of reliability can do the work previously done by notions of evidence or good reasons in distinguishing knowledge from merely true belief, and ***if*** naturalistic accounts are forthcoming of the concepts of *belief* and of *truth*, ***then*** a naturalistic account is possible of knowledge. That

[9] These are not equivalent characterizations: broadly naturalistic explanations need not restrict themselves to the language of physics. But for the purposes of the argument here, the differences don't make a difference.

at least this *inference* is good is almost universally taken not only to be true, but to be *obviously* true. I think, however, that it is *not* a good inference. When we understand properly the sense in which facts about the reliability of a mechanism can be objective, we will see that appeals to objective probability fall short of enabling fully naturalistic accounts of knowledge—even given the optimistic assumptions built into the premises of the inference. Seeing why this is so (in the next section of the chapter) provides the clues needed to formulate (in the final section) the lesson that we really ought to learn from the Founding Insight—what I will call the Implicit Insight of reliabilism.

V

The difficulty is a straightforward and familiar one, although I believe that its significance has not fully been appreciated. An objective probability can only be specified relative to a reference class. And in the full range of cognitive situations epistemological theories are obliged to address—by contrast to the carefully idealized situations described in artificially restricted vocabularies to which concepts of objective probability are applied in the special sciences—the world as it objectively is, apart from our subjective interests and concerns (paradigmatically, explanatory ones), does not in general privilege one of the competing universe of possible reference classes as the correct or appropriate one. Relative to a choice of reference class, we can make sense of the idea of objective probabilities, and so of objective facts about the reliability of various cognitive mechanisms or processes—facts specifiable in a naturalistic vocabulary. But the proper choice of reference class is not itself objectively determined by facts specifiable in a naturalistic vocabulary. So there is something left over.

The best way I know to make this point is by considering Alvin Goldman's barn facade example. This is perhaps ironic, because Goldman originally introduced the case twenty years ago in a classic paper that demolished the pretensions of then-dominant *causal* theories of knowledge, precisely in order to make room for the sort of reliabilist

alternatives that have held sway ever since.[10] While I do think this kind of example is decisive against causal theories of knowledge, in the context of aspirations to naturalize epistemology by appeal to considerations of reliability, it is a double-edged sword.

We are to imagine a physiologically normal perceiver, in standard conditions for visual perception (facing the object, in good light, no lenses or mirrors intervening, and so on) who is looking at a red barn. It looks like a red barn, he has seen many red barns before, and he is moved to say, and to believe, that there is a red barn in front of him. In fact, there is a red barn in front of him causing him perceptually to say and believe that. So his claim and his belief are true. He has the best reasons a perceiver could have for his belief: all the evidence he possesses confirms that it is a red barn, and that he can see that it is. Of central importance to Goldman's original purpose is that we may suppose that the causal chain linking the perceiver to the red barn in front of him is ideal; it is just as such causal chains should be in cases of genuine perceptual knowledge. (We may not know how to formulate conditions on such chains necessary or sufficient to qualify them as knowledge, but whatever they may be we are stipulating that those conditions are met in this case.) The perceiver has a true belief, has good reasons for that belief, and stands in the right causal relations to the object of his belief. Surely, one wants to say at this point, what he has in such a case is perceptual knowledge if anything is.

But things are less clear as we describe the case further, moving to facts *external* to the perceiver's beliefs, to his perceptual processing and to causal relations between the perceiver and what is perceived. For although the red barn our hero thinks he sees is indeed a red barn, it is, unbeknownst to him, located in Barn Facade County. There the local hobby is building incredibly realistic *trompe l'oeil* barn facades. In fact, our man is looking at the *only* real barn in the county—though there are 999 facades. These facades are so cunningly contrived that they are visually indistinguishable from actual barns. Were our subject (counterfactually) to be looking at one of the facades, he would form exactly the same beliefs he actually did about the real barn. That is, he would, falsely

[10] "Discrimination and Perceptual Knowledge" Journal of Philosophy, vol 73, no 20 (1976).

now, believe himself to be looking at an actual barn. It is just an accident that he happened on the one real barn.

The question is, does he know there is a red barn in front of him? A good case can be made that he does not. For though he has a true belief, it is only *accidentally* true. It is only true because he happened to stumble on the one real barn, out of a thousand apparent ones. This seems to be a case of exactly the sort that the third condition on knowledge, the one distinguishing it from merely accidentally true beliefs, was introduced to exclude. If that is right—and I think it is—then it shows that classical justificatory epistemological internalism is inadequate.[11] It also shows that appeal to the causal chain linking the believer to what his belief is about is not adequate to distinguish knowledge from merely accidentally true belief: the surprising conclusion Goldman was originally after. For the presence of barn facades in the vicinity—indeed, their local preponderance—not only does not affect the beliefs the candidate knower can appeal to as evidence for or reasons justifying his belief, it is *causally* irrelevant to the process by which that belief was formed.

Goldman's positive conclusion, of course, is that the difference that makes the epistemological difference in such cases is that in the circumstances in which the belief was actually formed—that is, in Barn Facade County—the subject is not a *reliable* perceiver of barns. Forming a belief as to whether something is a barn by looking at it is not, in that vicinity, a reliable belief-forming mechanism. What is special about this case is just that the circumstances that render unreliable here what elsewhere would be a reliable process are *external* to the subject's beliefs and to their connection to their causal antecedents. Goldman took a giant step here. Both the critical argument and the positive suggestion he drew from it—the combination I will call 'Goldman's Insight'—are epoch-making philosophical moves. But what is the exact significance of Goldman's reliabilist insight? Once we have rejected narrowly causal theories of the third condition on

[11] I say 'classical', because it is open to an internalist to deny that in such cases (and in all corresponding cases of the generic 'Twin Earth' type) the *internal* states are the same in the veridical and the nonveridical cases. *All* the two cases have in common is that the subject cannot tell them apart. But this fact need not be construed as sufficient to identify their contents. This is the option that McDowell pursues.

knowledge, and also classical justificatory internalist theories, what consequences should we draw from the demonstration of the positive bearing of external matters of reliability on assessments of knowledge? In particular, does Goldman's Insight support *naturalistic* ambitions in epistemology?

I think not. One of the happy features of Goldman's example is that it literalizes the metaphor of *boundaries* of reference classes. For suppose that Barn Facade County is one of a hundred counties in the state, all the rest of which eschew facades in favor of actual barns. Then, considered as an exercise of a differential responsive disposition within the *state*, rather than within the county, our subject's process of perceptual belief-formation may be quite reliable, and hence when it in fact yields correct beliefs, it may underwrite attributions of perceptual *knowledge.* But then, if the whole country, consisting of fifty larger states, shares the habits of Barn Facade County—so that over the whole country (excepting this one state) facades predominate by a large margin—then considered as a capacity exercised in the *country*, the very same capacity will count as quite *un*reliable, and hence as insufficient to underwrite attributions of knowledge. And then again, in the whole *world,* barns may outnumber facades by a large margin. So considered with respect to that reference class , the capacity would once again count as reliable. And so on. Do we need to know about the relative frequencies of barns and facades in the solar system or the galaxy in order to answer questions about the cognitive status of our subject's beliefs? On the other hand, if instead of looking at ever broader reference classes, we turn our attention to ever narrower ones, we end up with a reference class consisting simply of the actual exercise of the capacity in looking at a real barn. Within *that* reference class, the probability of arriving at a true belief is 1, since the unique belief arrived at in that situation is actually true. So with respect to the narrowest possible reference class, the belief-forming mechanism is maximally reliable.

Which is the correct reference class? Is the perceiver an objectively reliable identifier of barns or not? I submit that the facts as described do not determine an answer. Relative to each reference class there is a clear answer, but nothing in the way

the world is privileges one of those reference classes, and hence picks out one of those answers. An argument place remains to be filled in, and the way the world objectively is does not, by itself, fill it in. Put another way, the reliability of the belief-forming mechanism (and hence the status of its true products as states of knowledge) varies depending on how we describe the mechanism and the believer. Described as apparently perceiving this barn, he is reliable and knows there is a barn in front of him. Described as an apparent barn-perceiver in this county, he is not reliable and does not know there is a barn in front of him. Described as an apparent barn-perceiver in the state, he is again reliable and a knower, while described as an apparent barn-perceiver in the country as a whole he is not. And so on. All these descriptions are equally true of him. All are ways of specifying his location that can equally be expressed in purely naturalistic vocabulary. But these naturalistically statable facts yield different verdicts about the perceiver's reliability, and hence about his status as a knower. And no naturalistically statable facts pick out one or another of these descriptions as uniquely privileged or correct. So the naturalistically statable facts do not, according to epistemological reliabilism, settle whether or not the perceiver is a knower in the case described.

Now the case described is exceptional in many ways. Not every cognitive situation admits of descriptions in terms of nested, equally natural reference classes that generate alternating verdicts of reliability and unreliability. But I am not claiming that the idea of reliability is of no cognitive or epistemological significance. I am not denying Goldman's Insight. But situations with the structure of the barn facade example can arise, and they are counterexamples to the claim that reliabilism underwrites a naturalized epistemology—the mistaken idea that may be called the 'Naturalistic Blindspot' of reliabilism.

VI

How, then, *ought* we to understand the significance of considerations of reliability in epistemology? How can we properly acknowledge both the Founding Insight and

Goldman's Insight, while avoiding both the Conceptual and that Naturalistic Blindspot? And if not naturalism, what? *Super*naturalism? I think the key to answering these important questions is to see that, far from being opposed to considerations of what is a good reason for what, concern with reliability should itself be understood as concern with the goodness of a distinctive kind of *inference*. I will call this idea the 'Implicit Insight' of epistemological reliabilism.

Epistemology is usually thought of as the theory of knowledge. But epistemological theories in fact typically offer accounts of when it is proper to *attribute* knowledge: for instance, where there is justified true belief, or where true beliefs have resulted from reliable belief-forming processes. Now a theory of knowledge can take this form. The two might be related as formal to material mode, in Carnap's terminology; instead of asking what *X*'s are, we ask when the term 'X' is properly applied. But the two need not be versions of the same question. In the case of knowledge, I think they stand in a more complex relationship.

What is one doing in taking someone to have knowledge? The traditional tripartite response surely has the right form. To begin with, one is attributing some sort of *commitment*: a belief. For the reasons indicated above in connection with the Conceptual Blindspot, I think that being so committed must be understood as taking up a stance in an *inferentially* articulated network—that is, one in which one commitment carries with it various others as its inferential consequences, and rules out others that are incompatible. Only as occupying a position in such a network can it be understood as *propositionally* (and hence *conceptually*) contentful. Corresponding to the traditional justification condition on attributions of knowledge, we may say that not just any commitment will do. For it to be *knowledge* one is attributing, one must also take the commitment to be one the believer is in some sense *entitled* to. Mindful of the Founding Insight, we need not assume that the only way a believer can come to be entitled to a propositionally contentful is by being able to offer an inferential justification of it. Instead, entitlement may be attributed on the basis of an assessment of the reliability of

the process that resulted in the commitment's being undertaken. We'll return to look more closely at attributions of reliability, our final topic, just below.

So to take someone to know something one must do two things: attribute a certain kind of inferentially articulated *commitment*, and attribute a certain kind of *entitlement* to that commitment.[12] But not all beliefs to which the believer is entitled count as knowledge. One takes them so to qualify only where one takes them in addition to be *true*. What is it to do that? Taking a claim or belief to be true is not attributing an especially interesting and mysterious property to it; it is doing something else entirely. It is *endorsing* the claim oneself. Spurious metaphysical problems concerning the property of truth are what one gets if one *mis*understands what one is doing in *adopting* a stance oneself—undertaking a commitment—on the model of *describing, characterizing,* or *attributing* a property to someone *else*'s commitment. A corresponding mistake would be to think of making a promise, for instance that one would drive one's friend to the airport, as attributing a special sort of property to the proposition that one will drive one's friend to the airport—a property whose relation to one's own motivational structure will then cry out for explanation.

In calling what someone has 'knowledge' one is doing three things: *attributing* a *commitment* that is capable of serving both as premise and as conclusion of inferences relating it to other commitments, *attributing entitlement* to that commitment, and *undertaking* that same commitment oneself.[13] Doing this is adopting a complex, essentially *socially* articulated stance or position in the game of giving and asking for reasons. I won't attempt to develop or defend this way of understanding knowledge as a normative social status here—I have done so at length in *Making It Explicit*.[14] I have sketched it here because of the perspective it gives us on the role of attributions of *reliability* in securing entitlement to beliefs.

---

[12] For an argument that these two sorts of normative status are essential elements of any game of giving and asking for reasons, see Chapter VI below.

[13] Chapter V below explores some of the consequences of this social perspectival articulation of normative attitudes.

[14] [Harvard University Press, 1994], especially chapters 3, 4, and 5.

For suppose that, in the same spirit in which we just asked what one is *doing* in taking someone to be a knower, we ask what one is *doing* in taking someone to be a *reliable* former of noninferential beliefs about, say, red barns in front of him. To take someone to be a reliable reporter of red barns, under certain circumstances, is to take it that his reports of barns, in those circumstances, are likely to be *true*. According to the account just offered, to do that is to be inclined to *endorse* those reports oneself. And that means that what one is doing in taking someone to be reliable is endorsing a distinctive kind of *inference*: an inference, namely, from the *attribution* to another of a propositionally contentful commitment acquired under certain circumstances to the *endorsement* or *undertaking* oneself of a commitment with that same content. Inferences exhibiting this socially articulated structure are *reliability inferences*. Endorsing such an inference is just what being prepared to *rely* on someone else as an informant consists in: being willing to use *his* commitments as premises in one's *own* inferences (including practical ones).

The possibility of extracting information from the remarks of others is one of the main points of the practice of assertion, and of attributing beliefs to others. So reliability inferences play an absolutely central role in the game of giving and asking for reasons—indeed every bit as central as the closely related but distinguishable assessments of the *truth* of others' claims and beliefs. That concern with reliability is not opposed to concern with what is a reason for what, but actually a crucially important species of it, is what I want to call the Implicit Insight of reliabilism. Reliabilism deserves to be called a form of epistemological *externalism* because assessments of reliability (and hence of knowledge) can turn on considerations external to the reasons possessed by the candidate knower himself. In those cases, such assessments concern the reasons possessed by the *assessor* of knowledge, rather than by the *subject* of knowledge. The lesson I want to draw is that they should not therefore be seen as external to the game of giving and asking for reasons, nor to concern with what is a reason for what. Reliabilism points to

the fundamental *social* or *interpersonal* articulation of the practices of reason giving and reason assessing within which questions of who has knowledge arise.

A final dividend that this way of thinking about reliability pays is that it permits us to see what is really going on in the barn facade cases, and so how to take on board Goldman's insight. For the relativity to reference class of assessments of reliability (and hence of knowledge) that seemed so puzzling when viewed in a context that excluded concern with what is a reason for what fall naturally into place once we understand assessments of reliability as issues of what *inferences* to endorse. The different reference classes just correspond to different (true) collateral premises or auxiliary hypotheses that can be conjoined with the attribution of noninferentially acquired perceptual belief in order to extract inferential consequences the assessor of reliability (and knowledge) can use as premises in her *own* inferences. From the perceiver's report of a red barn and the premise that he is located in Barn Facade County, there is *not* a good inference to the conclusion that there is a red barn in front of him. From the perceiver's report and the premise that he is located in the *state*, there *is* a good inference to that conclusion. From the report and the premise that he is located in the *country*, there is not a good inference to that conclusion. And so on. All those collateral premises are true, so there are a number of candidate reliability inferences to be assessed. But there is no contradiction, because they are all *different* inferences. Nothing spooky or supernatural is going on—of course. The relativity to description that is threatening to an understanding of reliability and knowledge that ignores reason-giving, justification, and inference can be taken in stride once we see concern with reliability as arising in just such contexts. For we expect the goodness of inferences to be sensitive to differences how the items we are reasoning about are described. The intensionality of assessments of reliability is just a mark of their membership in the inferential order rather than the causal order. And we saw in the previous chapter that we should expect material inferences of this sort to be nonmonotonic.

To avoid the Conceptual Blindspot, one must appreciate the significance of specifically *inferential* articulation in distinguishing representations that qualify as *beliefs*, and hence as candidates for knowledge. To avoid the Naturalistic Blindspot, one must appreciate that concern with reliability is concern with a distinctive interpersonal *inferential* structure. Appreciating the role of inference in these explanatory contexts is grasping the Implicit Insight of reliabilism. It is what is required to conserve and extend both the Founding Insight, and Goldman's Insight, without being crippled by the difficulties into which they tempt us.