

## Chapter 9

### Epistemic Folkways and Scientific Epistemology

---

I

What is the mission of epistemology, and what is its proper methodology? Such meta-epistemological questions have been prominent in recent years, especially with the emergence of various brands of “naturalistic” epistemology. In this paper, I shall reformulate and expand upon my own meta-epistemological conception (most fully articulated in Goldman 1986), retaining many of its former ingredients while reconfiguring others. The discussion is by no means confined, though, to the meta-epistemological level. New substantive proposals will also be advanced and defended.

Let us begin, however, at the meta-epistemological level, by asking what role should be played in epistemology by our ordinary epistemic concepts and principles. By some philosophers’ lights, the sole mission of epistemology is to elucidate commonsense epistemic concepts and principles: concepts like knowledge, justification, and rationality, and principles associated with these concepts. By other philosophers’ lights, this is not even part of epistemology’s aim. Ordinary concepts and principles, the latter would argue, are fundamentally naive, unsystematic, and uninformed by important bodies of logic and/or mathematics. Ordinary principles and practices, for example, ignore or violate the probability calculus, which ought to be the cornerstone of epistemic rationality. Thus, on the second view, proper epistemology must neither *end* with naive principles of justification or rationality, nor even *begin* there.

My own stance on this issue lies somewhere between these extremes. To facilitate discussion, let us give a label to our commonsense epistemic concepts and norms; let us call them our *epistemic folkways*. In partial agreement with the first view sketched above, I would hold that *one* proper task of epistemology is to elucidate our epistemic folkways. Whatever else epistemology might proceed to do, it should at least have its roots in the concepts and practices of the folk. If these roots are utterly rejected and abandoned, by what rights would the new discipline call itself ‘epistemology’ at all? **It may well be desirable to reform or transcend our epistemic**

folkways, as the second of the views sketched above recommends. But it is essential to preserve continuity; and continuity can only be recognized if we have a satisfactory characterization of our epistemic folkways. Actually, even if one rejects the plea for continuity, a description of our epistemic folkways is in order. How would one know what to criticize, or what needs to be transcended, in the absence of such a description? So a first mission of epistemology is to describe or characterize our folkways.

Now a suitable description of these folk concepts, I believe, is likely to depend on insights from cognitive science. Indeed, identification of the semantic contours of many (if not all) concepts can profit from theoretical and empirical work in psychology and linguistics. For this reason, the task of describing or elucidating folk epistemology is a *scientific* task, at least a task that should be informed by relevant scientific research.

The second mission of epistemology, as suggested by the second view above, is the formulation of a more adequate, sound, or systematic set of epistemic norms, in some way(s) transcending our naive epistemic repertoire. How and why these folkways might be transcended, or improved upon, remains to be specified. This will partly depend on the contours of the commonsense standards that emerge from the first mission. On my view, epistemic concepts like knowledge and justification crucially invoke psychological faculties or processes. Our folk understanding, however, has a limited and tenuous grasp of the processes available to the cognitive agent. Thus, one important respect in which epistemic folkways should be transcended is by incorporating a more detailed and empirically based depiction of psychological mechanisms. Here too epistemology would seek assistance from cognitive science.

Since both missions of epistemology just delineated lean in important respects on the deliverances of science, specifically cognitive science, let us call our conception of epistemology *scientific epistemology*. Scientific epistemology, we have seen, has two branches: *descriptive* and *normative*. While **descriptive** scientific epistemology aims to describe our ordinary epistemic **assessments**, **normative** scientific epistemology continues the practice of **making** epistemic judgments, or formulating systematic principles for such **judgments**.<sup>1</sup> It is prepared to depart from our ordinary epistemic judgments, however, if and when that proves advisable. (This overall conception of epistemology closely parallels the conception of metaphysics articulated in chapters 2 and 3. The descriptive and normative branches of scientific epistemology are precise analogues of the descriptive and prescriptive branches of metaphysics, as conceptualized there.) In the remainder of this paper, I shall sketch and defend the particular forms of **descriptive and normative scientific epistemology that I favor**.

## II

Mainstream epistemology has concentrated much of its attention on two concepts (or terms): knowledge and justified belief. The preceding essay primarily illustrates the contributions that cognitive science can make to an understanding of the former; this essay focuses on the latter. We need not mark this concept exclusively by the phrase 'justified belief'. A family of phrases pick out roughly the same concept: 'well-founded belief', 'reasonable belief', 'belief based on good grounds', and so forth. I shall propose an account of this concept that is in the reliabilist tradition, but departs at a crucial juncture from other versions of reliabilism. My account has the same core idea as Ernest Sosa's *intellectual virtues* approach, but incorporates some distinctive features that improve its prospects.<sup>2</sup>

The basic approach is, roughly, to identify the concept of justified belief with the concept of belief obtained through the exercise of intellectual virtues (excellences). Beliefs acquired (or retained) through a chain of "virtuous" psychological processes qualify as justified; those acquired partly by cognitive "vices" are derogated as unjustified. This, as I say, is a *rough* account. To explain it more fully, I need to say things about the psychology of the epistemic evaluator, the possessor and deployer of the concept in question. At this stage in the development of semantical theory (which, in the future, may well be viewed as part of the "dark ages" of the subject), it is difficult to say just what the relationship is between the meaning or "content" of concepts and the form or structure of their mental representation. In the present case, however, I believe that an account of the form of representation can contribute to our understanding of the content, although I am unable to formulate these matters in a theoretically satisfying fashion.

The hypothesis I wish to advance is that the epistemic evaluator has a mentally stored set, or list, of cognitive virtues and vices. When asked to evaluate an actual or hypothetical case of belief, the evaluator considers the processes by which the belief was produced, and matches these against his list of virtues and vices. If the processes match virtues only, the belief is classified as justified. If the processes are matched partly with vices, the belief is categorized as unjustified. If a belief-forming scenario is described that features a process not on the evaluator's list of either virtues or vices, the belief may be categorized as neither justified nor unjustified, but simply *nonjustified*. Alternatively (and this alternative plays an important role in my story), the evaluator's judgment may depend on the (judged) *similarity* of the novel process to the stored virtues and vices. In other words, the "matches" in question need not be perfect.

**This proposal makes two important points of contact with going theories in the psychology of concepts. First, it has some affinity to the exemplar**

approach to concept representation (cf. Medin and Schaffer 1978; Smith and Medin 1981; Hintzman 1986). According to that approach, a concept is mentally represented by means of representations of its positive instances, or perhaps types of instances. For example, the representation of the concept *pants* might include a representation of a particular pair of faded blue jeans and/or a representation of the type *blue jeans*. Our approach to the concept of justification shares the spirit of this approach insofar as it posits a set of examples of virtues and vices, as opposed to a mere abstract characterization—e.g., a definition—of (intellectual) virtue or vice. A second affinity to the exemplar approach is in the appeal to a similarity, or matching, operation in the classification of new target cases. According to the exemplar approach, targets are categorized as a function of their similarity to the positive exemplars (and dissimilarity to the foils). Of course, similarity is invoked in many other approaches to concept deployment as well (see E. E. Smith 1990). This makes our account of justification consonant with the psychological literature generally, whether or not it meshes specifically with the exemplar approach.

Let us now see what this hypothesis predicts for a variety of cases. To apply it, we need to make some assumptions about the lists of virtues and vices that typical evaluators mentally store. I shall assume that the virtues include belief formation based on sight, hearing, memory, reasoning in certain “approved” ways, and so forth. The vices include intellectual processes like forming beliefs by guesswork, wishful thinking, and ignoring contrary evidence. *Why* these items are placed in their respective categories remains to be explained. As indicated, I plan to explain them by reference to reliability. Since the account will therefore be, at bottom, a reliabilist type of account, it is instructive to see how it fares when applied to well-known problem cases for standard versions of reliabilism.

Consider first the demon-world case. In a certain possible world, a Cartesian demon gives people deceptive visual experiences, which systematically lead to false beliefs. Are these vision-based beliefs justified? Intuitively, they are. The demon’s victims are presented with the same sorts of visual experiences that we are, and they use the same processes to produce corresponding beliefs. For most epistemic evaluators, this seems sufficient to induce the judgment that the victims’ beliefs are justified. Does our account predict this result? Certainly it does. The account predicts that an epistemic evaluator will match the victims’ vision-based processes to one (or more) of the items on his list of intellectual virtues, and therefore judge the victims’ beliefs to be justified.

Turn next to Laurence Bonjour’s (1985) cases in which hypothetical agents are assumed to possess a perfectly reliable clairvoyant faculty. Although these agents form their beliefs by this reliable faculty, Bonjour contends that the beliefs are not justified; and apparently most (philosoph-

ical) evaluators agree with that judgment. This result is not predicted by simple forms of reliabilism.<sup>3</sup> What does our present theory predict? Let us consider the four cases in two groups. In the first three cases (Samantha, Casper, and Maud), the agent has contrary evidence that he or she ignores. Samantha has a massive amount of apparently cogent evidence that the president is in Washington, but she nonetheless believes (through clairvoyance) that the president is in New York City. Casper and Maud each has large amounts of ostensibly cogent evidence that he/she has no reliable clairvoyant power, but they rely on such a power nonetheless. Here our theory predicts that the evaluator will match these agent’s belief-forming processes to the vice of ignoring contrary evidence. Since the processes include a vice, the beliefs will be judged to be unjustified.

Bonjour’s fourth case involves Norman, who has a reliable clairvoyant power but no reasons for or against the thesis that he possesses it. When he believes, through clairvoyance, that the president is in New York City, while possessing no (other) relevant evidence, how should this belief be judged? My own assessment is less clear in this case than the other three cases. I am tempted to say that Norman’s belief is *nonjustified*, not that it is thoroughly *unjustified*. (I construe *unjustified* as “having negative justificational status”, and *nonjustified* as “lacking positive justificational status”.) This result is also readily predicted by our theory. On the assumption that I (and other evaluators) do not have clairvoyance on my list of virtues, the theory allows the prediction that the belief would be judged neither justified nor unjustified, merely nonjustified. For those evaluators who would judge Norman’s belief to be *unjustified*, there is another possible explanation in terms of the theory. There is a class of putative faculties, including mental telepathy, ESP, telekinesis, and so forth that are scientifically disreputable. It is plausible that evaluators view any process of basing beliefs on the supposed deliverances of such faculties as vices. It is also plausible that these evaluators judge the process of basing one’s belief on clairvoyance to be *similar* to such vices. Thus, the theory would predict that they would view a belief acquired in this way as unjustified.<sup>4</sup>

Finally, consider Alvin Plantinga’s (1988) examples that feature disease-triggered or mind-malfunctioning processes. These include processes engendered by a brain tumor, radiation-caused processes, and the like. In each case Plantinga imagines that the process is reliable, but reports that we would not judge it to be justification conferring. My diagnosis follows the track outlined in the Norman case. At a minimum, the processes imagined by Plantinga fail to match any virtue on a typical evaluator’s list. So the beliefs are at least nonjustified. Furthermore, evaluators may have a prior representation of pathological processes as examples of cognitive vices. Plantinga’s cases might be judged (relevantly) similar to these vices, so that the beliefs they produce would be declared unjustified.

In some of Plantinga's cases, it is further supposed that the hypothetical agent possesses countervailing evidence against his belief, which he steadfastly ignores. As noted earlier, this added element would strengthen a judgment of unjustifiedness according to our theory, because ignoring contrary evidence is an intellectual vice. Once again, then, our theory's predictions conform with reported judgments.

Let us now turn to the question of how epistemic evaluators acquire their lists of virtues and vices. What is the basis for their classification? As already indicated, my answer invokes the notion of reliability. Belief-forming processes based on vision, hearing, memory, and ("good") reasoning are deemed virtuous because they (are deemed to) produce a high ratio of true beliefs. Processes like guessing, wishful thinking, and ignoring contrary evidence are deemed vicious because they (are deemed to) produce a low ratio of true beliefs.

We need not assume that each epistemic evaluator chooses his/her catalogue of virtues and vices by direct application of the reliability test. Epistemic evaluators may partly inherit their lists of virtues and vices from other speakers in the linguistic community. Nonetheless, the hypothesis is that the selection of virtues and vices rests, ultimately, on assessments of reliability.

It is not assumed, of course, that all speakers have the same lists of intellectual virtues and vices. They may have different opinions about the reliability of processes, and therefore differ in their respective lists.<sup>5</sup> Or they may belong to different subcultures in the linguistic community, which may differentially influence their lists. Philosophers sometimes seem to assume great uniformity in epistemic judgments. This assumption may stem from the fact that it is mostly the judgments of philosophers themselves that have been reported, and they are members of a fairly homogeneous subculture. A wider pool of "subjects" might reveal a much lower degree of uniformity. That would conform to the present theory, however, which permits individual differences in catalogues of virtues and vices, and hence in judgments of justifiedness.

If virtues and vices are selected on the basis of reliability and unreliability, respectively, why doesn't a hypothetical case introducing a novel reliable process induce an evaluator to add that process to his list of virtues, and declare the resulting belief justified? Why, for example, doesn't he add clairvoyance to his list of virtues, and rule Norman's beliefs to be justified?

I venture the following explanation. First, people seem to have a trait of *categorical conservatism*. They display a preference for "entrenched" categories, in Nelson Goodman's (1955) phraseology, and do not lightly supplement or revise their categorical schemes. An isolated single case is not enough. More specifically, merely imaginary cases do not exert much influence on categorical structures. People's cognitive systems are respon-

sive to live cases, not purely fictional ones. Philosophers encounter this when their students or nonphilosophers are unimpressed with science fiction-style counterexamples. Philosophers become impatient with this response because they presume that possible cases are on a par (for counterexample purposes) with actual ones. This phenomenon testifies, however, to a psychological propensity to take an invidious attitude toward purely imaginary cases.

To the philosopher, it seems both natural and inevitable to take hypothetical cases seriously, and if necessary to restrict one's conclusions about them to specified "possible worlds". Thus, the philosopher might be inclined to hold, "If reliability is the standard of intellectual virtue, shouldn't we say that clairvoyance is a virtue *in the possible worlds* of BonJour's examples, if not a virtue in general?" This is a natural thing for philosophers to say, given their schooling, but there is no evidence that this is how people naturally think about the matter. There is no evidence that "the folk" are inclined to relativize virtues and vices to this or that possible world.

I suspect that concerted investigation (not undertaken here) would uncover ample evidence of conservatism, specifically in the normative realm. In many traditional cultures, for example, loyalty to family and friends is treated as a cardinal virtue.<sup>6</sup> This view of loyalty tends to persist even through changes in social and organizational climate, which undermine the value of unqualified loyalty. Members of such cultures, I suspect, would continue to view personal loyalty as a virtue even in *hypothetical* cases where the trait has stipulated unfortunate consequences.

In a slightly different vein, it is common for both critics and advocates of reliabilism to call attention to the relativity of reliability to the domain or circumstances in which the process is used. The question is therefore raised, what is the relevant domain for judging the reliability of a process? A critic like John Pollock (1986, pp. 118–119), for example, observes that color vision is reliable on earth but unreliable in the universe at large. In determining the reliability of color vision, he asks, which domain should be invoked? Finding no satisfactory reply to this question, Pollock takes this as a serious difficulty for reliabilism. Similarly, Sosa (1988 and 1991) notes that an intellectual structure or disposition can be reliable with respect one field of propositions but unreliable with respect to another, and reliable in one environment but unreliable in another. He does not view this as a difficulty for reliabilism, but concludes that any talk of intellectual virtue must be relativized to field and environment.

Neither of these conclusions seems apt, however, for purposes of *description* of our epistemic folkways. It would be a mistake to suppose that ordinary epistemic evaluators are sensitive to these issues. It is likely—or at least plausible—that our ordinary apprehension of the intellectual vir-

tues is rough, unsystematic, and insensitive to any theoretical desirability of relativization to domain or environment. Thus, as long as we are engaged in the description of our epistemic folkways, it is no criticism of the account that it fails to explain what domain or environment is to be used. Nor is it appropriate for the account to introduce relativization where there is no evidence of relativization on the part of the folk.

Of course, we do need an explanatory story of how the folk arrive at their selected virtues and vices. And this presumably requires some reference to the domain in which reliability is judged. However, there may not be much more to the story than the fact that people determine reliability scores from the cases they personally “observe”. Alternatively, they *may* regard the observed cases as a sample from which they infer a truth ratio in some wider class of cases. It is doubtful, however, that they have any precise conception of the wider class. They probably don’t address this theoretical issue, and don’t do (or think) anything that commits them to any particular resolution of it. It would therefore be wrong to expect descriptive epistemology to be fully specific on this dimension.

A similar point holds for the question of process individuation. It is quite possible that the folk do not have highly principled methods for individuating cognitive processes, for “slicing up” virtues and vices. If that is right, it is a mistake to insist that descriptive epistemology uncover such methods. It is no flaw in reliabilism, considered as descriptive epistemology, that it fails to unearth them. It may well be desirable to develop sharper individuation principles for purposes of normative epistemology (a matter we shall address in section III). But the missions and requirements of descriptive and normative epistemology must be kept distinct.

This discussion has assumed throughout that the folk have lists of intellectual virtues and vices. What is the evidence for this? In the moral sphere ordinary language is rich in virtues terminology. By contrast, there are few common labels for intellectual virtues, and those that do exist—‘perceptiveness’, ‘thoroughness’, ‘insightfulness’, and so forth—are of limited value in the present context. I propose to identify the relevant intellectual virtues (at least those relevant to *justification*) with the belief-forming capacities, faculties, or processes that would be accepted as answers to the question “How does X know?”. In answer to this form of question, it is common to reply, “He saw it”, “He heard it”, “He remembers it”, “He infers it from such-and-such evidence”, and so forth. Thus, basing belief on seeing, hearing, memory, and (good) inference are in the collection of what the folk regard as intellectual virtues. Consider, for contrast, how anomalous it is to answer the question “How does X know?” with “By guesswork”, “By wishful thinking”, or “By ignoring contrary evidence”. This indicates that *these modes of belief formation—guessing, wishful thinking, ignoring contrary evidence—are standardly regarded as intellectual vices.*

They are not ways of obtaining knowledge, nor ways of obtaining justified belief.

Why appeal to “knowledge”-talk rather than “justification”-talk to identify the virtues? Because ‘know’ has a greater frequency of occurrence than ‘justified’, yet the two are closely related. Roughly, justified belief is belief acquired by means of the same sorts of capacities, faculties, or processes that yield knowledge in favorable circumstances (i.e., when the resulting belief is true and there are no Gettier complications, or no relevant alternatives).

To sum up the present theory, let me emphasize that it depicts justificational evaluation as involving two stages. The first stage features the acquisition by an evaluator of some set of intellectual virtues and vices. This is where reliability enters the picture. In the second stage, the evaluator applies his list of virtues and vices to decide the epistemic status of targeted beliefs. At this stage, there is no direct consideration of reliability.

There is an obvious analogy here to rule utilitarianism in the moral sphere. Another analogy worth mentioning is Saul Kripke’s (1980) theory of *reference-fixing*. According to Kripke, we can use one property to fix a reference to a certain entity, or type of entity; but once this reference has been fixed, that property may cease to play a role in identifying the entity across various possible worlds. For example, we can fix a reference to heat as the phenomenon that causes certain sensations in people. Once heat has been so picked out, this property is no longer needed, or relied upon, in identifying heat. A phenomenon can count as heat in another possible world where it doesn’t cause those sensations in people. Similarly, I am proposing, we initially use reliability as a test for intellectual quality (virtue or vice status). Once the quality of a faculty or process has been determined, however, it tends to retain that status in our thinking. At any rate, it isn’t reassessed each time we consider a fresh case, especially a purely imaginary and bizarre case like the demon world. Nor is quality relativized to each possible world or environment.

The present version of the virtues theory appears to be a successful variant of reliabilism, capable of accounting for most, if not all, of the most prominent counterexamples to earlier variants of reliabilism.<sup>7</sup> The present approach also makes an innovation in naturalistic epistemology. Whereas earlier naturalistic epistemologists have focused exclusively on the psychology of the epistemic agent, the present paper (along with the preceding essay) also highlights the psychology of the epistemic evaluator.

### III

Let us turn now to *normative scientific epistemology*. It was argued briefly in section I that *normative scientific epistemology should preserve continu-*

ity with our epistemic folkways. At a minimum, it should rest on the same types of evaluative criteria as those on which our commonsense epistemic evaluations rest. Recently, however, Stephen Stich (1990) has disputed this sort of claim. Stich contends that our epistemic folkways are quite idiosyncratic and should not be much heeded in a reformed epistemology. An example he uses to underline his claim of idiosyncrasy is the notion of justification as rendered by my “normal worlds” analysis in Goldman 1986. With hindsight, I would agree that that particular analysis makes our ordinary notion of justification look pretty idiosyncratic. But that was the fault of the analysis, not the analysandum. On the present rendering, it looks as if the folk notion of justification is keyed to dispositions to produce a high ratio of true beliefs in the actual world, not in “normal worlds”; and there is nothing idiosyncratic about that. Furthermore, there seem to be straightforward reasons for thinking that true belief is worthy of positive valuation, if only from a pragmatic point of view, which Stich also challenges. The pragmatic utility of true belief is best seen by focusing on a certain subclass of beliefs, viz., beliefs about one’s own *plans of action*. Clearly, true beliefs about which courses of action would accomplish one’s ends will help secure these ends better than false beliefs. Let proposition  $P =$  “Plan  $N$  will accomplish my ends” and proposition  $P' =$  “Plan  $N'$  will accomplish my ends”. If  $P$  is true and  $P'$  is false, I am best off believing the former and not believing the latter. My belief will guide my choice of a plan, and belief in the true proposition (but not the false one) will lead me to choose a plan that *will* accomplish my ends. Stich has other intriguing arguments that cannot be considered here, but it certainly appears that true belief is a perfectly sensible and stable value, not an idiosyncratic one.<sup>8</sup> Thus, I shall assume that normative scientific epistemology should follow in the footsteps of folk practice and use reliability (and other truth-linked standards) as a basis for epistemic evaluation.

If scientific epistemology retains the fundamental standard(s) of folk epistemic assessment, how might it diverge from our epistemic folkways? One possible divergence emerges from William Alston’s (1988) account of justification. Although generally sympathetic with reliabilism, Alston urges a kind of constraint not standardly imposed by reliabilism (at least not process reliabilism.) This is the requirement that the processes from which justified beliefs issue must have as their input, or basis, a state of *which the cognizer is aware* (or can easily become aware). Suppose that Alston is right about this as an account of our folk conception of justification. It may well be urged that this ingredient needn’t be retained in a scientifically sensitive epistemology. In particular, it may well be claimed that one thing to be learned from cognitive science is that only a small proportion of our cognitive processes operate on consciously accessible inputs. It could there-

fore be argued that a reformed conception of intellectually virtuous processes should dispense with the “accessibility” requirement.

Alston aside, the point of divergence I wish to examine concerns the psychological units that are chosen as virtues or vices. The lay epistemic evaluator uses casual, unsystematic, and largely introspective methods to carve out the mental faculties and processes responsible for belief formation and revision. Scientific epistemology, by contrast, would utilize the resources of cognitive science to devise a more subtle and sophisticated picture of the mechanisms of belief acquisition. I proceed now to illustrate how this project should be carried out.

An initial phase of the undertaking is to sharpen our conceptualization of the types of cognitive units that should be targets of epistemic evaluation. Lay people are pretty vague about the sorts of entities that qualify as intellectual virtues or vices. In my description of epistemic folkways, I have been deliberately indefinite about these entities, calling them variously “faculties”, “processes”, “mechanisms”, and the like. How should systematic epistemology improve on this score?

A first possibility, enshrined in the practice of historical philosophers, is to take the relevant units to be cognitive *faculties*. This might be translated into modern parlance as *modules*, except that this term has assumed a rather narrow, specialized meaning under Jerry Fodor’s (1983) influential treatment of modularity. A better translation might be (cognitive) *systems*, e.g., the visual system, long-term memory, and so forth. Such systems, however, are also suboptimal candidates for units of epistemic analysis. Many beliefs are the outputs of two or more systems working in tandem. For example, a belief consisting in the visual classification of an object (“That is a chair”) may involve matching some information in the visual system with a category stored in long-term memory. A preferable unit of analysis, then, might be a *process*, construed as the sort of entity depicted by familiar flow charts of cognitive activity. This sort of diagram depicts a sequence of operations (or sets of parallel operations), ultimately culminating in a belief-like output. Such a sequence may span several cognitive systems. This is the sort of entity I had in mind in previous publications (especially Goldman 1986) when I spoke of “cognitive processes”.

Even this sort of entity, however, is not a fully satisfactory unit of analysis. Visual classification, for example, may occur under a variety of degraded conditions. The stimulus may be viewed from an unusual orientation; it may be partly occluded, so that only certain of its parts are visible; and so forth. Obviously, these factors can make a big difference to the reliability of the classification process. Yet it is one and the same process that analyzes the stimulus data and comes to a perceptual “conclusion”. So the same process can have different degrees of reliability depending on a variety of parameter values. For purposes of epistemic assessment, it would

be instructive to identify the parameters and parameter values that are critically relevant to degrees of reliability. The virtues and vices might then be associated not with processes per se, but with processes operating *with specified parameter values*. Let me illustrate this idea in connection with visual perception.

Consider Irving Biederman's (1987, 1990) theory of object recognition, recognition-by-components (RBC). The core idea of Biederman's theory is that a common concrete object like a chair, a giraffe, or a mushroom is mentally represented as an arrangement of simple primitive volumes called *geons* (*geometrical ions*). These geons, or primitive "components" of objects, are typically symmetrical volumes lacking sharp concavities, such as blocks, cylinders, spheres, and wedges. A set of twenty-four types of geons can be differentiated on the basis of dichotomous or trichotomous contrasts of such attributes as curvature (straight versus curved), size variation (constant versus expanding), and symmetry (symmetrical versus asymmetrical). These twenty-four types of geons can then be combined by means of six relations (e.g., top-of, side-connected, larger-than, etc.) into various possible multiple-geon objects. For example, a cup can be represented as a cylindrical geon that is side-connected to a curved, handle-like geon, whereas a pail can be represented as the same two geons bearing a different relation: the curved, handle-like geon is at the top of the cylindrical geon.

Simplifying a bit, the RBC theory of object recognition posits five stages of processing. (1) In the first stage, low-level vision extracts edge characteristics, such as L's, Y-vertices, and arrows. (2) On the basis of these edge characteristics, viewpoint-independent attributes are detected, such as curved, straight, size-constant, size-expanding, etc. (3) In the next stage, selected geons and their relations are activated. (4) Geon activation leads to the activation of object models, that is, familiar models of simple types of objects, stored in long-term memory. (5) The perceived entity is then "matched" to one of these models, and thereby identified as an instance of that category or classification. (In this description of the five stages, all processing is assumed to proceed bottom-up, but in fact Biederman also allows for elements of top-down processing.)

Under what circumstances, or what parameter values, will such a sequence of processing stages lead to *correct*, or *accurate*, object identification? Biederman estimates that there are approximately 3,000 common basic-level, or entry-level, names in English for familiar concrete objects. However, people are probably familiar with approximately ten times that number of object models because, among other things, some entry-level terms (such as *lump* and *chair*) have several readily distinguishable object models. Thus, an estimate of the number of familiar object models would be on the order of 30,000.

Some of these object models are simple, requiring fewer than six components to appear complete; others are complex, requiring six to nine components to appear complete. Nonetheless, Biederman gives theoretical considerations and empirical results suggesting that an arrangement of only *two* or *three* geons almost always suffices to specify a simple object and even most complex ones. Consider the number of possible two-geon and three-geon objects. With twenty-four possible geons, Biederman says, the variations in relations can produce 186,624 possible two-geon objects. A third geon with its possible relations to another geon yields over 1.4 billion possible three-geon objects. Thus, if the 30,000 familiar object models were distributed homogeneously throughout the space of possible object models, Biederman reasons, an arrangement of two or three geons would almost always be sufficient to specify any object. Indeed, Biederman puts forward a *principle of geon recovery*: If an arrangement of two or three geons can be recovered from the image, objects can be quickly recognized even when they are occluded, rotated in depth, novel, extensively degraded, or lacking in customary detail, color, and texture.

The principle of three-geon sufficiency is supported by the following empirical results. An object such as an elephant or an airplane is complex, requiring six or more geons to appear complete. Nonetheless, when only three components were displayed (the others being occluded), subjects still made correct identifications in almost 80 percent of the nine-component objects and more than 90 percent of the six-component objects. Thus, the reliability conferred by just three geons and their relations is quite high. Although Biederman doesn't give data for recovery of just one or two geons of complex objects, presumably the reliability is much lower. Here we presumably have examples of parameter values—(1) number of components in the complete object, and (2) number of recovered components—that make a significant difference to reliability. The same process, understood as an instantiation of one and the same flow diagram, can have different levels of reliability depending on the values of the critical parameters in question. Biederman's work illustrates how research in cognitive science can identify both the relevant flow of activity and the crucial parameters. The quality (or "virtue") of a particular (token) process of belief-acquisition depends not only on the flow diagram that is instantiated, but on the parameter values instantiated in the specific tokening of the diagram.

Until now reliability has been my sole example of epistemic quality. But two other dimensions of epistemic quality—which also invoke truth or accuracy—should be added to our evaluative repertoire. These are *question-answering power* and *question-answering speed*. (These are certainly reflected in our epistemic folkways, though not well reflected in the concepts of knowledge or justification.) If a person asks himself a question,

such as “What kind of object is that?” or “What is the solution to this algebra problem?”, there are three possible outcomes: (A) he comes up with *no answer* (at least none that he believes), (B) he forms a belief in an answer which is *correct*, and (C) he forms a belief in an answer which is *incorrect*. Now reliability is the ratio of cases in category (B) to cases in categories (B) and (C), that is, the proportion of true beliefs to beliefs. Question-answering *power*, on the other hand, is the ratio of (B) cases to cases in categories (A), (B), and (C). Notice that it is possible for a system to be highly reliable but not very powerful. An object-recognition system that never yields outputs in category (C) is perfectly reliable, but it may not be very powerful, since most of its outputs could fall in (A) and only a few in (B). The human (visual) object-recognition system, by-contrast, is very powerful as well as quite reliable. In general, it is power and not just reliability that is an important epistemic desideratum in a cognitive system or process.

Speed introduces another epistemic desideratum beyond reliability and power. This is another dimension on which cognitive science can shed light. It might have been thought, for example, that correct identification of complex objects like an airplane or an elephant requires more time than simple objects such as a flashlight or a cup. In fact, there is no advantage for simple objects, as Biederman’s empirical studies indicate. This lack of advantage for simple objects could be explained by the geon theory in terms of parallel activation: geons are activated in parallel rather than through a serial trace of the contours of the object. Whereas more geons would require more processing time under a serial trace, this is not required under parallel activation.

Let us turn now from perception to learning, especially language learning. Learnability theory (Gold 1967; Osherson, Stob, and Weinstein 1985) uses a criterion of learning something like our notion of power, viz., the ability or inability of the learning process to arrive at a correct hypothesis after some fixed period of time. This is called *identification in the limit*. In language learning, it is assumed that the child is exposed to some information in the world, e.g., a set of sentences parents utter, and the learning task is to construct a hypothesis that correctly singles out the language being spoken. The child is presumed to have a learning strategy: an algorithm that generates a succession of hypotheses in response to accumulating evidence. What learning strategy might lead to success? *That* children learn their native language is evident to common sense. But *how* they learn it—what algorithm they possess that constitutes the requisite intellectual virtue—is only being revealed through research in cognitive science.

We may distinguish two types of evidence that a child might receive about its language (restricting attention to the language’s grammar): **positive evidence and negative evidence. Positive evidence refers to informa-**

tion about which strings of words *are* grammatical sentences in the language, and negative evidence refers to information about which strings of words *are not* grammatical sentences. Interestingly, it appears that children do not receive (much) negative evidence. The absence of negative evidence makes the learning task much harder. What algorithm might be in use that produces success in this situation?

An intriguing proposal is advanced by Robert Berwick (1986; cf. Pinker 1990). In the absence of negative evidence, the danger for a learning strategy is that it might hypothesize a language that is a superset of the correct language, i.e., one that includes all grammatical sentences of the target language plus some additional sentences as well. Without negative evidence, the child will be unable to learn that the “extra” sentences are incorrect, i.e., don’t belong to the target language. A solution is to avoid ever hypothesizing an overly general hypothesis. Hypotheses should be *ordered* in such a way that the child always guesses the narrowest possible hypothesis or language at each step. This is called the *subset principle*. Berwick finds evidence of this principle at work in a number of domains, including concepts, sound systems, and syntax. Here, surely, is a kind of intellectual disposition that is not dreamed of by the “folk”.

#### IV

We have been treating scientific epistemology from a purely reliabilist, or veritistic (truth-linked), vantage point. It should be stressed, however, that scientific epistemology can equally be pursued from other evaluative perspectives. You need not be a reliabilist to accept the proposed role of cognitive science in scientific epistemology. Let me illustrate this idea with the so-called *responsibilist* approach, which characterizes a justified or rational belief as one that is the product of epistemically responsible action (Kornblith 1983; Code 1987), or perhaps epistemically responsible processes (Talbot 1990). Actually, this conception of justification is approximated by my own *weak* conception of justification, as presented in chapter 7. Both depict a belief as justified as long as its acquisition is *blameless* or *nonculpable*. Given limited resources and limited information, a belief might be acquired nonculpably even though its generating processes are not virtuous according to the reliabilist criterion.

Let us start with a case of Hilary Kornblith. Kornblith argues that the justificational status of a belief does not depend exclusively on the *reasoning* process that produces that belief. Someone might reason perfectly well from the evidence he possesses, but fail to be epistemically responsible because he neglects to acquire certain further evidence. Kornblith gives the case of Jones, a headstrong young physicist eager to hear the praise of his colleagues. After Jones presents a paper, a senior colleague makes an

objection. Unable to tolerate criticism, Jones pays no attention to the objection. The criticism is devastating, but it makes no impact on Jones's beliefs because he does not even hear it. Jones's conduct is epistemically irresponsible. But his reasoning process from the evidence he actually possesses—which does not include the colleague's evidence—may be quite impeccable.

The general principle suggested by Kornblith's example seems to be something like this. Suppose that an agent (1) believes *P*, (2) does not believe *Q*, and (3) would be unjustified in believing *P* if he did believe *Q*. If, finally, he is *culpable* for failing to believe *Q* (for being ignorant of *Q*), then he is unjustified in believing *P*. In Kornblith's case, *P* is the physics thesis that Jones believes. *Q* consists in the criticisms of this thesis presented by Jones's senior colleague. Jones does not believe *Q*, but if he did believe *Q*, he would be unjustified in believing *P*. However, although Jones does not believe *Q*, he is culpable for failing to believe it (for being ignorant of these criticisms), because he *ought* to have paid attention to his colleague and acquired belief in *Q*. Therefore, Jones's belief in *P* is unjustified.

The provision that the agent be *culpable* for failing to believe *Q* is obviously critical to the principle in question. If the criticisms of Jones's thesis had never been presented within his hearing, nor published in any scientific journal, then Jones's ignorance of *Q* would not be culpable. And he might well be justified in believing *P*. But in Kornblith's version of the case, it seems clear that Jones *is* culpable for failing to believe *Q*, and that is why he is unjustified in believing *P*.

Under what circumstances is an agent culpable for failing to believe something? That is a difficult question. In a general discussion of culpable ignorance, Holly Smith (1983) gives an example of a doctor who exposes an infant to unnecessarily high concentrations of oxygen and thereby causes severe eye damage. Suppose that the latest issue of the doctor's medical journal describes a study establishing this relationship, but the doctor hasn't read this journal. Presumably his ignorance of the relationship would be culpable; he *should* have read his journal. But suppose that the study had appeared in an obscure journal to which he does not subscribe, or had only appeared one day prior to this particular treatment. Is he still culpable for failing to have read the study by the time of the treatment?

Smith categorizes her example of the doctor as a case of *deficient investigation*. The question is (both for morals and for epistemology), What amounts and kinds of investigation are, in general, sufficient or deficient? We may distinguish two types of investigation: (1) investigation into the physical world (including statements that have been made by other agents), and (2) investigation into the agent's own storehouse of information, lodged in long-term memory. Investigation of the second sort is particularly rele-

vant to questions about the role of cognitive science, so I shall concentrate here on this topic. Actually, the term 'investigation' is not wholly apt when it comes to long-term memory. But it is adequate as a provisional delineation of the territory.

To illustrate the primary problem that concerns me here, I shall consider two examples drawn from the work of Amos Tversky and Daniel Kahneman. The first example pertains to their study of the "conjunction fallacy" (Tversky and Kahneman 1983). Suppose that a subject assigns a higher probability to a conjunction like "Linda is a bank teller and is active in the feminist movement" than to one of its own conjuncts, "Linda is a bank teller". According to the standard probability calculus, no conjunction can have a higher probability than one of its conjuncts. Let us assume that the standard probability calculus is, in some sense, "right". Does it follow that a person is irrational, or unjustified, to make probability assignments that violate this calculus? This is subject to dispute. One might argue that it does not follow, in general, from the fact that *M* is an arbitrary mathematical truth, that anyone who believes something contrary to *M* is ipso facto irrational or unjustified. After all, mathematical facts are not all so transparent that it would be a mark of irrationality (or the like) to fail to believe any of them. However, let us set this issue aside. Let us imagine the case of a subject who has studied probability theory and learned the conjunction rule in particular. Let us further suppose that this subject would retract at least one of his two probability assignments if he recognized that they violate the conjunction rule. (This is by no means true of all subjects that Tversky and Kahneman studied.) Nonetheless, our imagined subject fails to think of the conjunction rule in connection with the Linda example. Shall we say that the failure to recover the conjunction rule from long-term memory is a *culpable omission*, one that makes his maintenance of his probability judgments unjustified? Is this like the example of Jones who culpably fails to learn of his senior colleague's criticism? Or is it a case of *nonculpable nonrecovery* of a relevant fact, a fact that is, in some sense "within reach", but legitimately goes unnoticed?

This raises questions about when a failure to recover or activate something from long-term memory is culpable, and that is precisely a problem that invites detailed reflection on mechanisms of memory retrieval. This is not a matter to which epistemologists have devoted much attention, partly because little has been known about memory retrieval until fairly recently. But now that cognitive science has at least the beginnings of an understanding of this phenomenon, normative epistemology should give careful attention to that research. Of course, we cannot expect the issue of culpability to be resolved directly by empirical facts about cognitive mechanisms. Such facts are certainly relevant, however.

The main way that retrieval from memory works is by *content addressing* (cf. Potter 1990). Content addressing means starting retrieval with part of the content of the to-be-remembered material, which provides an “address” to the place in memory where identical or similar material is located. Once a match has been made, related information laid down by previously encoded associations will be retrieved, such as the name or appearance of the object. For example, if you are asked to think of a kind of bird that is yellow, a location in memory is addressed where “yellow bird” is located. “Yellow bird” has previously been associated with “canary”, so the latter information is retrieved. Note, however, that there are some kinds of information that cannot be used as a retrieval address, although the information is in memory. For example, what word for a family relationship (e.g., *grandmother*) ends in *w*? Because you have probably never encoded that piece of information explicitly, you may have trouble thinking of the word (hint: not *niece*). Although it is easy to move from the word in question (*nephew*) to “word for a family relationship ending in *w*”, it is not easy to move in the opposite direction.

Many subjects who are given the Linda example presumably have not established any prior association between such pairs of propositions (“Linda is a bank teller and is active in the feminist movement” and “Linda is a bank teller”) and the conjunction rule. Furthermore, in some versions of the experiment, subjects are not given these propositions adjacent to one another. So it may not occur to the subject even to *compare* the two probability judgments, although an explicit comparison would be more likely to address a location in memory that contains an association with the conjunction rule. In short, it is not surprising, given the nature of memory retrieval, that the material provided in the specified task does not automatically yield retrieval of the conjunction rule for the typical subject.

Should the subject deliberately search memory for facts that might retrieve the conjunction rule? Is omission of such deliberate search a culpable omission? Perhaps, but how much deliberate attention or effort ought to be devoted to this task? (Bear in mind that agents typically have numerous intellectual tasks on their agendas, which vie for attentional resources.) Furthermore, what form of search is obligatory? Should memory be probed with the question, “Is there any rule of probability theory that my (tentative) probability judgments violate?” This is a plausible search probe for someone who has already been struck by a thought of the conjunction rule and its possible violation, or whose prior experiences with probability experiments make him suspicious. But for someone who has not already retrieved the conjunction rule, or who has not had experiences with probability experiments that alert him to such “traps”, what reason is there to be on the lookout for violations of the probability calculus? It is

highly questionable, then, that the subject engaged in “deficient investigation” in failing to probe memory with the indicated question.

Obviously, principles of culpable retrieval failure are not easy to come by. Any principles meriting our endorsement would have to be sensitive to facts about memory mechanisms.

A similar point can be illustrated in connection with the so-called *availability heuristic*, which was formulated by Tversky and Kahneman (1973) and explored by Richard Nisbett and Lee Ross (1980). A cognizer uses the availability heuristic when he estimates the frequency of items in a category by the instances he can *bring to mind* through memory retrieval, imagination, or perception. The trouble with this heuristic, as the abovementioned researchers indicate, is that the instances one brings to mind are not necessarily well correlated with objective frequency. Various *biases* may produce discrepancies: biases in initial sampling, biases in attention, or biases in manner of encoding or storing the category instances.

Consider some examples provided by Nisbett and Ross: one hypothetical example and one actual experimental result. (1) (Hypothetical example) An Indiana businessman believes that a disproportionate number of Hoosiers are famous. This is partly because of a bias in initial exposure, but also because he is more likely to notice and remember when the national media identify a famous person as a Hoosier. (2) (Actual experiment) A group of subjects consistently errs in judging the relative frequency of words with *R* in first position versus words with *R* in third position. This is an artifact of how words are encoded in memory (as already illustrated in connection with *nephew*). We don’t normally code words by their third letters, and hence words having *R* in the third position are less “available” (from memory) than words beginning with *R*. But comparative availability is not a reliable indicator of actual frequency.

Nisbett and Ross (p. 23) view these uses of the availability heuristic as normative errors. “An indiscriminate use of the availability heuristic,” they write, “clearly can lead people into serious judgmental errors.” They grant, though, that in many contexts perceptual salience, memorability, and imaginability may be relatively unbiased and well correlated with true frequency or causal significance. They conclude: “The normative status of using the availability heuristic ... thus depend[s] on the judgmental domain and context. People are not, of course, totally unaware that simple availability criteria must sometimes be discounted. For example, few people who were asked to estimate the relative number of moles versus cats in their neighborhood would conclude ‘there must be more cats because I’ve seen several of them but I’ve never seen a mole.’ Nevertheless, as this book documents, people often fail to distinguish between legitimate and superficially similar, but illegitimate, uses of the availability heuristic.”

We can certainly agree with Nisbett and Ross that the availability heuristic can often lead to incorrect estimates of frequency. But does it follow that uses of the heuristic are often *illegitimate* in a sense that implies the epistemic *culpability* of the users? One might retort, "These cognizers are using all the evidence that they possess, at least *consciously* possess. Why are they irresponsible if they extrapolate from this evidence?" The objection apparently lurking in Nisbett and Ross's minds is that these cognizers *should* be aware that they are using a systematically biased heuristic. This is a piece of evidence that they *ought* to recognize. And their failure to recognize it, and/or their failure to take it into account, makes their judgmental performance culpable. Nisbett and Ross's invocation of the cat/mole example makes the point particularly clear. If someone can appreciate that the relative number of cats and moles *he has seen* is not a reliable indicator of the relative number of cats and moles in the neighborhood, surely he can be expected to appreciate that the relative number of famous Hoosiers *he can think of* is not a reliable indicator of the proportion of famous people who are Hoosiers!

Is it so clear that people *ought* to be able to appreciate the biased nature of their inference pattern in the cases in question? Perhaps it seems transparent in the mole and Hoosier cases; but consider the letter R example. What is (implicitly) being demanded here of the cognizer? First, he must perform a feat of meta-cognitive analysis: he must recognize that he is inferring the relative proportion of the two types of English words from his own constructed samples of these types. Second, he must notice that his construction of these samples depends on the way words are encoded in memory. Finally, he must realize that this implies a bias in ease of retrieval. All these points may seem obvious in hindsight, once pointed out by researchers in the field. But how straightforward or obvious are these matters if they haven't already been pointed out to the subject? Of course, we currently have no "metric" of straightforwardness or obviousness. That is precisely the sort of thing we need, however, to render judgments of culpability in this domain. We need a systematic account of how difficult it is, starting from certain information and preoccupations, to generate and apprehend the truth of certain relevant hypotheses. Such an account clearly hinges on an account of the inferential and hypothesis-generating strategies that are natural to human beings. This is just the kind of thing that cognitive science is, in principle, capable of delivering. So epistemology must work hand in hand with the science of the mind. The issues here are not purely scientific, however. Judgments of justifiedness and unjustifiedness, on the responsibility conception, require assessments of culpability and nonculpability. **Weighing principles for judgments of culpability is a matter for philosophical attention. (One question, for example, is how much epistemic culpability depends on voluntariness.)** Thus, a mix of phi-

losophy and psychology is needed to produce acceptable principles of justifiedness.

### Notes

I wish to thank Tom Senor, Holly Smith, and participants in a conference at Rice University for helpful comments on earlier versions of this paper.

1. Normative scientific epistemology corresponds to what I elsewhere call *epistemics* (see Goldman 1986). Although epistemics is not restricted to the assessment of *psychological* processes, that is the topic of the present paper. So we are here dealing with what I call *primary epistemics*.
2. Sosa's approach is spelled out most fully in Sosa 1985, 1988, and 1991.
3. My own previous formulations of reliabilism have not been so simple. Both "What Is Justified Belief?" (chapter 6 of this volume) and *Epistemology and Cognition* (Goldman 1986) had provisions—e.g., the non-undermining provision of *Epistemology and Cognition*—that could help accommodate Bonjour's examples. It is not entirely clear, however, how well these qualifications succeeded with the Norman case, described below.
4. Tom Senor presented the following example to his philosophy class at the University of Arkansas. Norman is working at his desk when out of the blue he is hit (via clairvoyance) with a very distinct and vivid impression of the president at the Empire State Building. The image is phenomenally distinct from a regular visual impression but is in some respects similar and of roughly equal force. The experience is so overwhelming that Norman just can't help but form the belief that the president is in New York. About half of Senor's class judged that in this case Norman justifiably believes that the president is in New York. Senor points out, in commenting on this paper, that their judgments are readily explained by the present account, because the description of the clairvoyance process makes it sufficiently similar to vision to be easily "matched" to that virtue.
5. Since some of these opinions may be true and others false, people's lists of virtues and vices may have varying degrees of accuracy. The "real" status of a trait as a virtue or vice is independent of people's opinions about that trait. However, since the enterprise of descriptive epistemology is to describe and explain evaluators' judgments, we need to advert to the traits they *believe* to be virtues or vices, i.e., the ones on their mental lists.
6. Thanks to Holly Smith for this example. She cites Riding 1989 (chap. 6) for relevant discussion.
7. It should be noted that this theory of justification is intended to capture what I call in chapter 7 the *strong* conception of justification. The complementary conception of *weak* justification will receive attention in section IV of this essay.
8. For further discussion of Stich, see Goldman 1991.