

DAVID HENDERSON and TERENCE E. HORGAN

PRACTICING SAFE EPISTEMOLOGY\*

(Received in revised form 24 November 1999)

Reliabilists have argued that the important evaluative epistemic concept of being justified in holding a belief, at least to the extent that that concept is associated with knowledge, is best understood as concerned with the objective appropriateness of the processes by which a given belief is generated and sustained. They hold that a belief is justified only when it is fostered by processes that are reliable in the believer's actual world.<sup>1</sup> Of course, reliabilists typically recognize other concepts of justification – subjective notions – which are given a noncompeting sort of epistemic legitimacy. However, they focus on the epistemically central notion of “strong justification,” and have come to settle on this familiar reliabilist analysis, supposing that it pretty much exhausts what there is to say about “objective justification.”

The straightforward reliabilist analysis of objective justification has contributed to epistemological understanding. However, there is yet clarification and perspective to be gotten by recognizing further epistemically valuable features that are distinct from, but related to, reliability. These additional epistemically valued features are “objective” in much the sense that reliability is. We here develop a way of thinking about one such epistemic value and suggest that it may also have an important role in our thinking about an agent being objectively justified in holding a given belief. Like the reliability of generating processes, the feature we characterize is epistemically valuable in view of the epistemic interest in the production of true belief systems.

Our suggestion is that understanding these further epistemic values will allow one to better appreciate both the place for reliability in epistemic evaluations and its limits. The limits do *not* indicate that reliabilists have been mistaken in thinking that the



reliability in the agent's world of fostering processes is importantly related to an agent's being strongly justified. Yet, they do indicate that the reliabilist position is improved by recognizing how related but distinct evaluative concerns also feature in the objective appropriateness of processes. Seeing this much allows us to appreciate how a common set of misgivings regarding the traditional reliabilist analysis can be accommodated within a perspective that remains focused on objective (rather than subjective) features of processing, and on features closely related to reliability.

#### 1. THE RECEIVED ACCOUNT OF STRONG JUSTIFICATION

It has become common to distinguish objective justification (sometimes called strong justification, warrant, or some brand of positive epistemic status) from subjective justification. The basic idea is that justification of the stripe that is associated with having knowledge must have to do with objectively appropriate processing, and not simply processing that conforms to whatever epistemic norms the agent happens to have internalized. Goldman (1992a), for example, reflects on evaluations we might make when looking at beliefs formed by agents in an epistemically benighted society. In such cases, he observes, we can feel pulled in two directions. On the one hand, reflective and conscientious agents might do their best and have subjectively powerful reasons for their beliefs; and we would want to say that such agents are justified in their beliefs. On the other hand, such agents might be doing their best and nevertheless be employing processes that are objectively inappropriate to the central epistemic goal of producing true belief-systems; and we would want to say that those agents are not justified in their beliefs. Both evaluations have deep roots in our epistemic tradition. In view of the deep difference in what anchors these evaluations, we then also should recognize that they reflect distinct uses of 'justification'. So doing, we disambiguate our talk of justification.

So, we have this notion of objective justification, or warrant, that centers on the notions of objectively appropriate processing.<sup>2</sup> In the most general terms, objective appropriateness is a matter of what would be conducive to the central epistemic end of the production of true belief-systems.<sup>3</sup> It is perhaps worth emphasizing at this point

that our central epistemic interest, that in terms of which objective appropriateness of processing must be understood, is not simply in the production of true beliefs. Rather, it is an interest in the production of true beliefs *systems*.<sup>4</sup> Loosing sight of the interest in the systematicity of our beliefs can lead to distorted understandings of our epistemic standards. Reliability, and the value we characterize here – robustness – are largely truth-focused standards. Were one to lose sight of the interest in systematicity, one might readily think of such values as pushing us in the direction of highly conservative epistemic practice. But, such practices would not conduce to the production of interesting systematicity of beliefs. Such features of processing must not be thought to make for epistemic appropriateness of themselves. Rather, echoing Goldman (1986), they must be thought of as values that, along with the “power” of processes for the production of beliefs (and the power for the production of systematicity within beliefs) make for the objective appropriateness of processing. We will have occasion to reiterate such themes later in this work, but it will be good for the reader to keep the present point, and the general view of objective appropriateness, in mind throughout this paper.

In any case, objective appropriateness is a matter of what would be conducive to the central epistemic end of the production of true belief-systems. More detailed models of objectively appropriate processing would be models whose realization by agents in their worlds would ensure the effectiveness of their cognition insofar as this is possible.<sup>5</sup> Even when a person is rationally pursuing the epistemic end in light of his or her understanding of the tendencies of various cognitive processes, that person may yet diverge from objectively appropriate processing. While doing what is appropriate from the first-person point of view, people may rely on hallucinogen-induced visions, on books handed down over centuries from visionaries or mystics, or even on what their mommies and daddies told them. But, such processes are not really objectively effective, and are not objectively appropriate.

What, in slightly less general terms, makes for objectively appropriate processing? On the received reliabilist account, what objectively conduces to our central epistemic end is basically truth-conduciveness or reliability. This is surely part of the story; at least

part of what can contribute to a cognitive process being objectively appropriate is its being reliable. But here we argue that there are further truth-related features of processes that properly enter into our epistemic evaluation. Reliability is but one component of what makes for objective appropriateness, and turns out not even to function as a necessary condition for objective justification.

What is truth-conducive varies across possible worlds. What is truth-conducive for us in our world would not be truth-conducive for agents in other worlds. So arises within the received reliabilist account the issue of what we should make of the justification (or lack of justification) of agents in very different worlds – agents who might, for example, be employing processes that would be truth-conducive in our world but not in theirs. Reliability is particularly directly tied to the furthering of our central epistemic end, and thus to the objective appropriateness of processes. But, for the agents in any given possible world, it is reliability *in that world* that is intimately connected with success in pursuing the epistemic end. *Ceteris paribus*, an agent's processes are epistemically improved in the relevant sense to the extent that the agent comes to employ processes with greater reliability in *that agent's* world. Thus the objective appropriateness of processes, and their objective justification, are a matter of the processes being truth-conducive in the agent's world. As Goldman (1992b) put the point, 'objective justification' is a nonrigid designator.

We grant that the above lines of thought characterize part of the story about the standards for objectively appropriate epistemic processing (and for objective justification). Our interest here is in recovering additional objectivist epistemic values that also serve as standards for appropriate processing and objective justification. The objective appropriateness of processing will be found not to be wholly a function of the reliability of those processes in the agent's world. In section 2, we will set out some basis for thinking that accounts wholly in terms of reliability are inadequate. In section 3, we elucidate a second objective feature of processing – robustness – that can be recognized as epistemically valuable. Robustness conduces to the end of producing true belief-systems, and its objective importance reflects significant elements of the epistemic situation. In section 4, we show that thinking in terms of

robustness allows us to deal with certain problems facing an account of the objective appropriateness of processes wholly in terms of reliability. We suggest in section 5 that, in order to do justice to considered epistemic evaluations, one ultimately needs a multi-dimensional understanding of objective epistemic appropriateness. However, the case for the multi-dimensional understanding is not simply that it accords with “intuitions.” Rather more importantly, we will see that there is a deep sense in these intuitions – they point to objective features of processes that are valuable in terms of the central epistemic value.

## 2. MISGIVING ABOUT RELIABILISM

There are familiar objections and counter-examples to reliabilism. For example, one’s reactions to hypothetical clairvoyants who form beliefs without (at first) any reason to think their own belief-forming processes reliable raise doubts regarding the supposed sufficiency of reliability for strong justification,<sup>6</sup> and generally about the intimacy of the connection between reliability and justification. However, Goldman (1992b) has articulated reasons for being suspicious of the force of many counter-examples. In effect, Goldman distinguishes between our evolving conceptualization and our concept.<sup>7</sup> Our judgments would seem to be largely controlled by conceptualizations, and thus need not reflect something deep and conceptually ensured about what it is to be justified. Goldman provides a plausible starting place in thinking about the structure of our conceptualization, suggesting that our conceptualization may commonly take the form of lists of (paradigmatic) approved and unapproved processes. Our basis for these lists may be slow in effecting a modification of our lists, and our use of these lists may be somewhat insensitive to counter-factual imaginations. That is, our conceptualization may be historically conditioned by concerns for reliability – concerns that are central to the concept – while the lists have come to be somewhat inflexibly internalized in most of us.

Goldman’s points in (1992b), together with care not to slip between objective and subjective notions of justification, do blunt the force of the standard counter-examples somewhat. Of themselves, such counter-examples should not be taken as decisive.

But, neither are they philosophically pointless. Rather, an adequate philosophical account will reflectively come to terms with these objections, dismissing them only on the basis of a considered understanding of the nature of the workings of our concept of objective justification. Such an account can allow one to explain why such apparent counter-examples can seem telling, when they are not. Thus, Goldman's strategy is philosophically appropriate. However, in what follows, we will suggest that better sense can be made of certain apparent counter-examples to straightforward reliabilism by taking them to be also real counter-examples. The counter-example developed here is closely related to what Sosa (1991) has dubbed the "new evil demon problem." It does not point to a wholesale repudiation of reliabilism, but to a recognition of features of processing that are closely related to reliability and that play a coordinate role in determining the epistemic appropriateness or inappropriateness of processing. This counter-example points us toward a refined reflective account of objective justification.

Our counter-example involves agents in a possible world of that fairly extreme sort characterized in some classical epistemological mythology: a demon-world. To prepare for this central thought experiment, it will be helpful to distinguish between two sorts of hypothetical scenarios that might be taken to characterize demon-worlds.

On the one hand, there is the scenario envisioned in Descartes' *Meditations*, one in which a malicious and powerful being seeks to defeat our epistemic project at every turn. Were the being really powerful and devoted – were the being really good at epistemic evil – then *whatever cognitive processes we were to employ*, the being would *adapt* so as to frustrate us. Call such a world a *classical demon-world*, or a *flexible-demon-world*. With due respect for Descartes, we must conclude that the fundamental epistemic good really could not be advanced in such a possible world. No matter what processes were employed on the input received, an agent in such a world would be doomed to abject failure.

On the other hand, there is a milder form of skeptical scenario sometimes entertained recently. One in which a brain is placed in a nutrient bath, and hooked up to a supercomputer that has been programmed to provide it systematically false input.<sup>8</sup> (As it is

commonly put, the computer takes into account the brain's output, and gives it just the input it would receive were it walking down the street, playing tennis, or so on for various standard life activities.) It is common to see this scenario as the modern, high-tech, analog to the classical demon-world. However, there is this important difference: by hypothesis, the computer is programmed to provide a certain class of misleading input. At least as normally described, the computer does not so change the input it gives the agent as to frustrate that agent no matter what *epistemic procedure* the agent employs. (While the input that the computer gives to the agent is conceived as tailored to decisions of the agent to walk one direction rather than another, to lean on one post rather than another, or to go one concert rather than another, this input is typically not thought to be so tailored as to anticipate the agent's reasoning processes and ensure that these also go wrong by giving the agent whatever input would lead just such reasoning awry.) Accordingly, we might call such worlds *rigid-demon-worlds*. Notably, depending on the details, the systematic falsity of the input may then allow for a systematic correction within the agent's cognitive processing.

What seems common across the various demon-world scenarios is that the appearances had by individuals in these epistemically possible worlds would be radically deceiving in undetectable (or very nearly undetectable) ways. However, if this formulation is to characterize both types of demon-worlds and begin to distinguish them from non-demon-worlds, we must understand "appearances" in a particular way. The appearances must not be simply taken as perceptual beliefs.<sup>9</sup> After all, perceptual beliefs are deeply colored by any given epistemic agent's fleshed out belief system – by the rich set of theories and associated ways of "seeing" things that a given individual has developed. But, supposing that an individual has then developed a very false belief system, that individual's perceptual beliefs could well be radically deceiving. Further, given the way in which belief systems can inform epistemic practice, there may be no way that such an individual can detect the deceiving character of the resulting perceptual beliefs. But, such a scenario does not constitute a demon-world, although it is an unfortunate epistemic situation. If we are to distinguish the particular classes of demon-worlds from these other unfortunate situations, appearances must be something

more rudimentary than perceptual beliefs, something “skinner” and more pervasive.

So, appearances are skinnier than many perceptual beliefs, something shared by varying perceptual beliefs, as these latter are differently colored by varying belief systems. Examples of what we have in mind are fairly easy to cite. We all seem disposed to “see” enduring three-dimensional objects – although how these are “perceived” will be colored by a rich background of beliefs. This can be witnessed even in small infants. They will give passing attention to spots moving behind and between a series of screens in a standard fashion, becoming bored and moving on. However, when the moving spots are so arranged as to give the appearance of discontinuity, seemingly to frustrate the “expectations” for enduring objects, infants will remain attentive longer, as if attempting to discern what might be going on. Of course, what we “perceive” as going on is richer than just “seeing” some enduring three-dimensional object yonder. Thus, where one person may “perceive” people traveling with a light at some distance, another may “perceive” spirits or witchcraft traveling at night (Evans-Pritchard 1937, p. 33–4), but in both cases there is the appearance of an enduring object. Also, both the envatted brain and ourselves in our normal world may “perceive” a table, and in this case we again have the appearances of an enduring object situated in such and such a fashion with respect to ourselves. Also, we seem to be so set up as to “perceive” persons – although what exactly persons are may be cultural informed (Geertz 1983). Yet, at some level, it seems fair to say that appearances are shared. Presumably, the brain in the vat has person-appearances. Our appearances of persons and enduring objects are, on the whole, not radically deceiving, the poor envatted brain’s would be.

The misgiving developed here has to do with what a straightforward reliabilist would need to say about agents in classical demon-worlds. It is clear that, in such worlds, agents have no prospect of getting to a true belief-system. In effect, there simply are no possible truth-conducive processes for an agent to employ in such a world – at least none for the generation of *a posteriori* beliefs. This leads the straightforward reliabilist to counter-intuitive results. On such a view, the reliability of generating/sustaining processes is

at least a necessary condition (and perhaps almost a sufficient condition) for objective propriety of processing, for objective justification. *It follows that agents in a classical demon-world could not be justified in any of their (a posteriori) beliefs – no matter what processing they were to employ. But, most of us judge that not all processes are equally objectively bad or inappropriate for agents in this world.*<sup>10</sup> Suppose, for example, that one agent, Constance, has taken note of where her observations have seemed untrustworthy in the past, and discounts certain observations accordingly. Suppose Constance only generalizes when samples are large enough for statistical confidence at some high level, and then only when the samples are either random or characterized by a diversity that seems to match the distribution of likely causal features in the population. Suppose that another agent, Faith, engages in the most extravagant flights of wishful thinking, believing what she wants and editing her observations and generalizations to suit. Now, suppose that Constance and Faith are in a classical demon-world. It seems extremely implausible to think that, because they are in such a world, Constance and Faith are equally lacking in strong justification for their beliefs. Surely, not all possible agents in a classical demon-world need be in the same sorry boat with respect to the objective appropriateness of their epistemic processes. (We should emphasize that the epistemic difference between Constance and Faith is not to be understood in terms of weak, or subjective justification. These agents may be equally conscientious in applying their respective epistemic norms. The relevant difference is that Constance's practice – and thus the norms that guide that practice – are *objectively* more appropriate than Faith's.)

Admittedly, in making these judgments, one could be relying on inflexible list-esque conceptualizations, as Goldman might suggest. But we think that there is a better understanding – it is that: (a) there are multiple features of cognitive processes that contribute to their objective epistemic appropriateness, and (b) unlike reliability, some of these help make for appropriateness independently of the world the agent happens to be in. The feature we will recommend as epistemically valuable here is not proposed simply to avoid counter-examples; rather, it is to be recognized as valuable for much the

same reasons that lead us to value reliability – each objectively conduces to furthering the central epistemic end.

Reliabilists might seek to soften the tone of their judgments of agents in demon-worlds. For example, a reliabilist might insist that, in worlds where there are no reliable processes possible, and thus no possibility of objectively appropriate processes and of objectively justified beliefs, it seems best to say that processes are neither appropriate nor inappropriate. The reliabilist might write of all processes in a classical demon-world as “nonappropriate,” and of all beliefs as “nonjustified.” While this has a kinder, more sympathetic, ring to it, it really does not respond to the core objection – it does not provide a basis for distinguishing between persons, or between processes, in classical demon-worlds. The core of the objection turns on the intuition that there can be differences in the objective appropriateness of processes of agents in such worlds. It is that the reliabilist has missed something of epistemological importance, not that the reliabilist has been too harsh in characterizing a homogeneous class of possible epistemic cases. The solution is not an epistemological form of “political correctness.” While all agents in a classical demon-world would be “epistemically disadvantaged,” we cannot leave the matter there. Some agents may employ processes that are epistemically appropriate, while others may fail to do so.

Of course, if these claims are to be defensible, one must be able to make clear sense of the idea that processes may have features that contribute to their being epistemically appropriate independent of the world the possessors happen to be in. This can be done. As we will show, the key is to keep in mind that uncertainty regarding the world in which one epistemically labors is characteristic of agents’ epistemic situation. It is a ubiquitous fact of epistemic life. Accordingly, at least some of what is objectively appropriate, will have to do with how epistemic agents can manage in the context of their uncertainty.

### 3. EXPENDING THE RANGE OF OBJECTIVE EPISTEMIC VALUES: ROBUSTNESS

Given the classic epistemological end of producing true belief-systems, there are multiple objective features of cognitive processes

that are objectively valuable. That is, there are multiple objective features of processes that conduce to the production of true belief-systems. Of course, reliability in the particular epistemically relevant possible world that the agent occupies is commonly and correctly recognized to be such a feature. In this section, *robustness of reliability*, or simply *robustness*, is shown to be another. Adding this feature to the list of appropriateness-contributing features enriches epistemological thought in important ways.

### 3.1. *Epistemically Relevant Possible Worlds*

Our characterization of robustness will make use of the notion of an epistemically relevant possible world – or (to put it more simply) an epistemically possible world. Let us explain this notion. There is an intuitive sense in which having experience with roughly the character of one's common, everyday, experience is compatible with a very wide range of possible worlds – an epistemic agent could have experience very much like the experience that we common epistemic agents have and yet be in very different worlds. Some of these worlds would be ones in which agents would be correct in taking much of their experience “at face value,” although what that value is may be also vary somewhat with theories. Others would be ones in which agents would have those experiences, but would be largely misled by them. This might conceivably happen in two ways. In one, the experiences would be misleading because of the coloring of appearance by socialization into a belief system.<sup>11</sup> The second is the more radical sense in which experience can be misleading: it would be a matter of even the appearances (with whatever coloring) being misleading – as would obtain were we brains in vats or were we set upon by an evil demon, for example. The set of epistemically relevant possible worlds runs this gamut of worlds in which agents would have appearances of the character of our everyday experience.

### 3.2. *Robustness and Its Epistemic Value*

Briefly, robustness of reliability may be characterized as *truth-conducivity in a very wide set of epistemically relevant possible worlds*. Recognizing the epistemic value of robustness will turn on recognizing that our epistemic endeavor must be undertaken in the

face of uncertainty, our fallible understanding regarding the world in which we are situated. Certain features of processes are objectively valuable to us fallible agents, and robustness is prominent. It should be emphasized that the value of robustness is parallel to the value of reliability: both are derivative from the central epistemic value of producing true belief-systems – which is taken as a given. We seek to produce or foster true belief-systems; given this central epistemic end, and given uncertainty, robustness turns out to be an objectively valuable feature of processes (something that is valuable, given the understood ends, whether particular agents recognize it or not).

Because epistemic agents, qua epistemic agents, have a fallible understanding of which epistemically possible world is the actual world, they have a clear interest in using methods that are reliable in a fairly wide set of epistemically relevant possible worlds – wide enough that it is safe to employ such methods or processes, despite uncertainty about which possible world is the actual world. Their fallibility is an epistemic fact of life that conditions what is epistemically valuable; it gives rise to the objective importance of epistemic prudence – or safe epistemology.

The present point parallels Cherniak's. Cherniak (1986) has urged that our finitude is characteristic of our epistemic situation, and that this fact has generally been under-appreciated. He shows that our finitude deeply conditions what sorts of processes are objectively appropriate (see also Henderson 1994a, 1994b). In much the same way, our uncertainty regarding the world in which we epistemically labor deeply conditions what processes are epistemically appropriate. To come to terms with the epistemic significance of this uncertainty, it is necessary to investigate how one can effectively pursue the production of true belief-system in the face of this uncertainty. The features of processes that help here are components of objective appropriateness; and here robustness is central.

Again, robustness is truth-conducivity in a very wide set of epistemically relevant possible worlds. Somewhat more precisely, it may be characterized as reliability in a wide set of epistemically relevant worlds other than those extreme classical demon worlds in which no method is reliable. One might say it is the property of a process being reliably reliable: reliably (that is across a very wide set of possible worlds in which reliable processes might be had)

such a process is reliable (that is produces mostly truth beliefs in the relevant world). It was noted above that one sort of epistemically possible world comprises those extreme hypothetical scenarios – the demon-worlds. In effect, there are no truth-conducive processes available to agents in such worlds. Now, for purposes of judging the robustness of reliability of a given process, it seems pointless to be concerned with how that process would fare in such worlds, for all processes would essentially fare the same way there, and one reasonably ignores such extremes when gauging dispositional features.

Compare judgments concerning the reliability of automobiles.<sup>12</sup> There are two senses in which we speak of automobiles as “reliable.” One has striking parallels with reliability in the agent’s world; the other with robustness. One may say that an individual’s car is reliable (or “dependable”), given that individual’s situation. Thus, we might say that a graduate student’s old car had proven reliable transportation in the mild climate in which that student was situated. We might also readily admit that that car would not be reliable were it employed in some more extreme climate. The sort of reliability/dependability at issue here corresponds to the reliability of the epistemological literature – reliability in an environment. The dependability of an auto is an objective feature of autos that varies across environments – in some environments, an auto will tend to yield transportation when it is turned to, while in other environments it may commonly fail. Similarly, a cognitive process would objectively tend to produce true belief-systems in some epistemically possible worlds, and not in others.

In contrast, there is a notion of reliability of automobiles that parallels the robustness of cognitive processes: we speak of an auto being reliable (or “durable”) when we note that it would prove reliable in a significant range of environments. Our graduate student’s aged auto may well not be reliable in this sense. To find out which cars are reliable, we do well to consult *Consumer Reports*, as their survey of readers reports the reliability-in-environment experience of readers who presumably represent a significant range of environments. However, there are possible environments in which no auto functions well – proximity to volcanic eruptions, or to detonations of thermonuclear devices, for example. We, and *Consumer Reports*,

ignore these cases as too extreme to matter. For similar reasons, the above formulation simply excludes demon-worlds from consideration in determining whether a process is robustly reliable. The standard for robustness of cognitive processes, like the standard for the reliability/durability of automobiles, is a matter of their working within a significant range of situations where some process (or auto) might work. With this standard understood, robustness of processes is, like reliability/durability of autos, an objective feature. Against the background of the central interest in transportation, reliability/durability is a desirable feature of autos, and against the background of the central interest of epistemology, robustness is a desirable feature of cognitive processes.

Robustness comes in degrees. All robust processes are reliable (or would “work”) in a very wide set of possible non-demon-worlds. But some would work in a more extensive set than others.<sup>13</sup> The ideal case of robustness, perfect robustness, would be reliability in all non-demon-worlds. A perfectly robust process would be wonderful, but this would be a lot to ask for. We know of no process capable of generating interesting, extensive, belief-systems that would measure up to such a standard. (And, because the central epistemic interest is in the production of such belief-systems, the value of robustness must be balanced with the value of the “power” and “system-conducivity” of our processes. Accordingly, we ultimately settle for moderate robustness.) Moderately robust processes would fail to be reliable in some non-demon-worlds. Yet they would be reliable in a very extensive set of non-demon-worlds. There are (presumably vague) limits to the worlds in which a process may fail and still count as moderately robust. But for our purposes here, we need not determine just how wide a set is extensive enough.

By way of illustration, consider various inductive processes. Inductive processes that are insensitive to possible sources of sample bias, and that do not include mechanisms for avoiding such bias, may yet be reliable in certain worlds – notably those in which populations are homogeneous. That is, if the same causal factors operate uniformly on the members of those populations which are the focus of inductions, then however we choose the sample we study, we will be looking at a subset that reflects the causal features operating throughout the larger population. So, even without sens-

itivity to sample bias, we will manage to look upon representative samples – no thanks to us. On the other hand, such a process would not be truth-conducive in a world where there is heterogeneity in the populations of interest. There, sensitivity to possible sample bias, and thus the ability to draw and work from representative samples, is requisite for reliability. So, inductive processes with this characteristic are reliable in a wider set of possible worlds than are those lacking it. Presumably, such processes might qualify as moderately robust; whereas those lacking it, while reliable in a select set of particularly benign worlds, are not robust. Of course, even inductive processes with sensitivity to possible sample bias would fail to be reliable in some epistemically relevant possible worlds. Notably, classical demon-worlds provide one set of worlds in which they are not reliable. We ignore these extreme scenarios when gauging the robustness of these methods. Still, there are less extreme epistemologically possible worlds – worlds that we do not ignore when gauging robustness – in which these processes fail to be reliable. For example, Humean recalcitrant worlds, that is worlds in which hitherto dependable regularities occasionally change in unprojectible fashion, provide one context in which even sensitivity to sample bias does not yield reliability. Thus these processes are not perfectly robust, although they may qualify as moderately robust.

Obviously, robust processes help meet the needs that we have as fallible agents – providing us a feature that at least partially compensates for our uncertainty regarding which world is the actual world, and thus regarding what processes are reliable in our world. There are several helpful and complementary perspectives from which to elucidate the epistemic value of robustness. We present two here.

To begin with, think of agents as shaping their own cognitive processes with the central epistemic end in view.<sup>14</sup> To some extent, the cognitive processes that agents employ are informed by, or conditioned by, their understanding of what processes are truth-conducive.<sup>15</sup> Whatever cognitive processes are employed, these will be truth-conducive in some worlds and not in others. However, just what this (the actual) world is like, and thus just what processes would be truth-conducive here, is not something about which we

are, or should be, certain. Agents must press on in the face of this uncertainty. In this choice situation, robustness is desirable.

Suppose that an epistemic agent employs a process would be truth-conducive only in a narrow range of epistemically possible worlds, where that range includes what that agent takes the world to be like. The agent thus employs a non-robust process that is otherwise appropriate to the world as understood. Now, were that agent to be mistaken about the world, the agent could easily be employing an unreliable process. *Employing such non-robust processes in the face of uncertainty about which epistemically possible world is the actual world is risky, it is to court epistemic failure.* On the other hand, suppose an epistemic agent employs a process that is robust, and suppose that this agent is yet mistaken regarding which possible world is the actual world. Nevertheless, it is relatively likely that that agent's world is among the extensive set of possible worlds in which the robust process that are employed would be truth-conducive. *Employing such robust processes is thus less risky.* In employing robust procedures, one is prudently employing processes that may well be truth-conducive, even if one is mistaken (or inattentive, or simply ignorant) concerning the world and what processes are truth-conducive in it. *In employing robust processes, we thus allow ourselves an epistemic margin for error.* Thereby, we practice safe epistemology.<sup>16</sup>

There is a more generic line of thought (a second elucidation of the value of robustness) that parallels the above line without presupposing that the cognitive processes of agents are, in any significant sense, conditioned by their understandings of what the world is like and what processes are reliable. As noted above, *if* their processes are so conditioned, then agents are at risk of using an unreliable process to the extent that their processes are not robust. But, suppose instead that their processes are not conditioned by their understandings of what the world is like. This could happen either due to lack of (implicit or explicit) beliefs regarding their world and the attendant reliability of processes or due to the absence of (either accessible or inaccessible) mechanisms conditioning processes to such beliefs. Again, to whatever extent their processes are not robust, such agents run a commensurate risk of employing unreliable processes. The generic point begins with the recognition that there are a range of

ways that the world might be compatible with appearances having much the character that they do – there are, that is, multiple epistemically possible worlds. Processes will vary in their reliability (or unreliability) across these worlds. Were an agent’s processes to be conditioned by an infallible understanding of the world, then obviously there is no risk of that agent’s processes being unreliable. But, this is simply not in the cards. To employ a process that is not conditioned to an infallible understanding of which epistemically possible world is the actual world is to run *some risk* of using an unreliable process. It is to employ a process that would be reliable in some epistemically possible worlds, and not in others, and for the process used not to be conditioned by infallible information so as to ensure reliability in the actual world. Agents fail to so condition, or “tailor,” their processes whenever they have and must draw on fallible understandings of the world – this was the subject of the first perspective on the value of robustness. Now we add that agents, *also* fail to so condition their processes whenever their processes are simply not conditioned by their understandings. All cognitive agents fall under one of these conditions. Put simply, all epistemic agents run some risk here (for they have fallible understandings, and may, to boot, be employing processes that are not even conditioned by such understanding).<sup>17</sup> The pivotal issue has to do with the degree of risk, with risk management. If one’s processes are perfectly robust, the risk is literally minimized. This is probably too much to ask for. *If one’s processes are moderately robust, the risk is moderated, and this is epistemically desirable. If, however, one’s processes are non-robust, their use is unacceptably risky from an epistemic point of view.*

### 3.3. *Epistemic Safety*

In view of these considerations, it seems natural to think of reliability and robustness as two forms of epistemic safety, both epistemically valuable, although one has not been widely recognized. Let us mention some implications. To begin with it is helpful to notice that talk of “safety” can reflect a diversity of conceptions. The most generic conception of safety is that of an item that can be employed with a minimum (or at least a relatively or acceptably low level) of risk. We can discern two more specific conceptions of safety, paral-

leling the two desiderata of reliability and robustness. An informed advisor bent on helping agents in their methodical pursuit of truth would be concerned with both.

First, there is what we might term *local safety*. This is a kind of brute, actual-world safety that comes with the likelihood that things will work out well, from the point of view of given ends, in the agent's environment or world. When the given end is the central epistemic end of fostering true belief-systems, local safety comes with reliability of processes in the agent's actual world.<sup>18</sup> Let us return to our earlier automotive analogies. When the goal is transportation (on demand), and when the climate is temperate and the terrain undemanding, then many old jalopies would be locally safe. In contrast, when the climate is extremely hot or cold, or when the terrain is rugged, fewer vehicles would be safely relied on. When the goal is daily transport with few injuries, and when the traffic density is low, the roads good, and there are few obstacles to run into, then even the classic Volkswagen Beetle, or the Ford Pinto, are locally safe autos. In many contexts, these autos are not locally safe. Were the traffic-density high, and were many (survivalist) locals to drive military-surplus armored vehicles, even the generally safe Volvo and Mercedes would not be locally safe. Similarly, given the central epistemic goal, when the world has only homogeneous populations, inductive processes that are not sensitive to sampling bias will yet be reliable and thus locally safe. On the other hand, worlds with less homogeneous populations will be ones in which inductive processes will only be locally safe when they are sensitive to the representativeness of samples.

Second, there is what we might call *general safety*. This is a kind of safety that turns on the likelihood that things will work out well from the point of view of given ends in a significant range of epistemically relevant possible worlds. When the given end is that of fostering true belief-systems, general safety comes with robustness. Again there are automotive analogies. While failing in certain extreme environments, Volkos and Mercedes are said to facilitate injury avoidance in a fairly wide range of environments, and might then qualify as generally safe autos. Inductive processes that include sensitivity to possible sample bias thereby avoid ways in which we might be led to false generalizations in the worlds with

nonhomogeneous populations. Of course, they will fail in worlds with special nonhomogeneous populations – say where hitherto dependable regularities occasionally just change in unprojectible Humean fashion – say tomorrow. But, such fairly extreme worlds comprise a relatively small set.

The notion of general safety serves to highlight central elements making for the objective value of robustness: a process with robustness can be employed with a view to the production and maintenance of true belief systems with a minimum (or, with moderate robustness, a relatively low level) of risks (of falling by producing false beliefs), where this risk is gauged in terms of ranges of epistemically relevant worlds that reflect the uncertainty characteristic of epistemic agents.

We should emphasize here that, because uncertainty regarding the epistemic playing field is a characteristic of epistemic life, a fact of epistemic life, and because robustness provides a needed prophylactic for the resulting epistemic risks, robustness becomes a pervasive epistemic value. It is valuable to all epistemic agents, simply qua epistemic agents. The point will be pursued further in section 5.

The epistemic importance of robustness is reflected at various points in our epistemological tradition. One that has served as an inspiration to us is the pragmatic justification of induction, which can readily be recast in terms of a concern for robustness. In the pragmatic justification of induction, one sought to show that induction would work (lead to true general beliefs), if any method would work. The idea was to provide an epistemic vindication of induction, without presupposing it. Of course, without presupposing induction, the vindication would need to proceed in radical ignorance or uncertainty regarding the world in which the processes were to be applied. For example, one could not presuppose anything about the degree of regularity in the world. The epistemic vindication then involved showing that, were there enough regularity for some method at all to work, then induction would work in that world. So, without knowing whether there were sufficient regularity for any method to work, we could engage in our pursuit of truth by employing induction. If there were not sufficient regularity, then our pursuit would end in failure – but we would not have foregone any alternative method that

would have been more fruitful. We would not have used “the wrong” epistemic method. On the other hand, if there is enough regularity, we will be using a truth-conducive method.<sup>19</sup>

One can understand the pragmatic justification of induction as an attempted demonstration that induction has robustness, working in a very wide class of non-demon-worlds. The pragmatic justification of induction supposes that we have observational input from the world that is to be trusted; thus, for purposes of the argument, it is supposed that we are in a non-demon-world. The issue it then addresses is whether induction provides a method that will take us further – beyond true particular beliefs to true systems of *general* beliefs. It seeks to show that there is a method for the generation of general beliefs (induction) that will work in any world in which there is some method for generating general beliefs that will work. Supposing that worlds with enough regularity for some method or other to work constitute a wide set of the non-demon-worlds, this would be to show robustness.<sup>20</sup>

#### 4. AN APPLICATION OF ROBUSTNESS TO AN IMPORTANT TEST CASE

Not everyone in a demon-world need be objectively epistemically equal. One should be able to distinguish between processes even in that context, finding some objectively appropriate and others not (or at least finding some more objectively appropriate than others). Such were the judgments that motivated our misgivings in section 2. Of course, such “intuitive” judgments can be mistaken. But, before we jump to such a conclusion, we should cast about for grounds that both comport with our general understanding of objective appropriateness and that would provide the basis for these intuitive judgments. In the last section, we argued that our general understanding of objective appropriateness leads us to take robustness of reliability as one feature that contributes to the objective appropriateness of processes. In this section, we discuss how robustness goes a long way to enabling us to account for our judgments about demon-world scenarios. On the basis of these points, it then would seem reasonable to conclude that such common intuitive judgments are informed by an inarticulate good epistemic sense –

that we are there responding to an appreciation of the objective epistemic value of robustness. In effect, intuitive judgments here have been more nuanced than much of the systematic epistemic literature at this point.

One plausible reconstruction of intuitive reasoning about agents in demon-worlds is as follows. We begin with the recognition that reliability of processes is not in the cards for any agent in such a world. It is simply not an epistemic value that can be realized there. We are thus led to judge that reliability provides an inappropriate dimension on which to evaluate processes in such a world, as it fails to provide for interesting distinctions. We then think in terms of other dimensions, other epistemic values, that might make for interesting differences; robustness is prominent.

The agent either did or did not practice safe epistemology – implanting cognitive processes that would work in any (or at least a wide range of) the epistemically relevant possible worlds in which there is some process that would work. Suppose that the agent used processes that should work (as an objective matter) as long as one were not so unlucky as to have “landed” in a demon-world (unknowingly, of course). As we argued in the preceding section, this feature transcends particular worlds and contributes to a process’s being objectively appropriate. We think that it looms large enough as a contributor to objective appropriateness that agents in demon worlds can be justified from the point of view of using objectively robust processes, even if not reliable processes.

So, our suggestion is that robustness comes to dominate thinking about demon-world scenarios as one reasons as follows: although an agent there cannot be employing reliable processes, such an agent might yet be justified – in the sense of using processes that are objectively appropriate – by virtue of employing robustly reliable processes. The importance of robustness in this connection seems in order because, when agents are employing such processes, their failures cannot be traced to their processes. Suppose that an agent has prudently employed robustly reliable processes. Such an individual would not only have done everything subjectively that could be done in the epistemic endeavor, *but also everything that objectively could be done in the epistemic endeavor*. Such agents would be employing processes that minimize their risk in the face of uncer-

tainty regarding the world in which they labor.<sup>21</sup> They would have done what is appropriate, and all that is objectively appropriate, in the face of the uncertainty that is characteristic of the epistemic situation. Had they not had the misfortune of landing in such an extremely inhospitable epistemic world, their processes would have been reliable.<sup>22</sup> Thus, their inevitable failure cannot be traced to them. There would be nothing about these agents that objectively contributed to their failure in any significant way. (We might then say that they bear no “objective responsibility” for their inevitable failure, or that their failure was not due to a fault in them.) Their processes are not part of the problem, and would serve as a solution in a significant range of worlds where some process could serve as a solution to the epistemic need. Put most simply, we should conclude: damn fine agents, damn lousy world. Since robustness provides a world-independent dimension making for objective appropriateness, we can add: damn fine processes, damn lousy world.<sup>23</sup>

##### 5. AN EMERGING MULTI-DIMENSIONAL ACCOUNT OF THE OBJECTIVE APPROPRIATENESS OF PROCESSING

We have argued that, in addition to reliability in the agent’s world, there are other very general features of processes that are epistemically valuable and that can contribute to a process being objectively appropriate. Robustness of reliability is another such feature – one that has been ill appreciated in the epistemological literature. Like reliability, its epistemic value is related to, one might say derivative from, the central epistemic value of the production of systems of true beliefs. (Of course, along with robustness and reliability, the power of processes and their contribution to systematicity of beliefs also must be kept in view.) We have traced the value of robustness to the uncertainty characteristic of the epistemic situation – it provides a measure of epistemic safety in the face of that pervasive uncertainty. We have shown how attention to robustness allows the avoidance of certain counter-intuitive results of straight reliabilism – doing so within the framework of objectivist understandings of epistemological warrant.

Still, some readers will have questions regarding the importance of robustness. Is it really a general epistemic value – in the fashion of reliability? Or does it rather come into play only, or primarily, in certain special, and extreme, contexts? It might be tempting to think that the robustness condition applies only when the reliability condition could have no application, only where there are no reliable processes to be had. But, while this might seem an open option when reflecting just on how robustness came to the fore almost by default in the classical demon-world case, it ignores a central point in our presentation. We have argued that uncertainty regarding the world in which we epistemically labor is itself a pervasive feature of the epistemic situation, that robustness is the indicated countermeasure, and that robustness consequently is an objectively valuable feature for epistemic agents generally. *Robustness is a pervasive epistemic value, because uncertainty is a pervasive fact of epistemic life.*

Reliabilists commonly recognize that reliability is, at best, a necessary, but not sufficient, condition for appropriateness of processing. For example, some insist that an appropriate process must also itself arise in a way that is conducive to truth-conducivity (Goldman 1992a). In imposing this additional requirement, a reliabilist comes very close to recognizing what we are after in the robustness condition. If agents generate their epistemic processes in an inherently risky fashion, then, if they happen to employ reliable processes because the world just happens to be one of the limited number of ways that make their processes reliable, *they are the undeserving beneficiaries of dumb luck*. Now, the central idea behind attention to meta-reliability seems to be an *aversion to an epistemic role for dumb luck*. The reliability of belief-spawning processes is not sufficient for the objective propriety of a belief, and this is shown when we reflect on cases where that reliability is significantly a matter of dumb luck on the part of the agent. The beneficiary of dumb luck is using processes that are not objectively appropriate – that are *too risky* from the epistemic point of view. The epistemic safety that comes with the use of robustly reliable belief-producing (and maintaining) processes minimizes the place for dumb luck in our epistemic success. In view of the uncertainty

characteristic of the epistemic situation, the use of robust reliable processes is epistemically desirable.<sup>24</sup>

So robustness and reliability generally function as two coordinate conditions on appropriateness. Each condition seems generally applicable, at least where satisfiable by any process at all. In the classical demon-world, reliability is not really satisfiable, thus it has little force or application there – robustness dominates. But in a wide range of epistemically possible worlds, both are satisfiable.<sup>25</sup>

Finally, a note on the application of our coordinate values of reliability and robustness. It is well known that the demand for reliability tends towards a kind of conservatism in the production of beliefs. Commonly, one refines a process so as to increase its reliability by foreclosing ways in which the unrefined processes would have gone wrong by producing false beliefs in the agent's environment. The refined processes are then more cautious, less free in their production of beliefs. Famously, caution can be taken too far. Descartes' professed method in *Meditation I* (as opposed to his practice in the *Meditations*) serves as an example – had he followed it, he would have produced no (or almost no) beliefs. Relibilists readily admit that trade-offs are in order. After a minimal level of reliability is provided for, one must trade some reliability for productiveness of processes. After all, the goal is to produce systems of true beliefs (and not just to insure against the production of false beliefs). Such considerations led Goldman (1986) to distinguish three epistemic values: speed and power, as well as reliability. We would add conduciveness to the systematicity of our beliefs as yet another epistemic value.

These familiar points apply, *mutatis mutandis*, to the epistemic value of robustness. Robust processes may be said to be relatively cautious when compared to alternatives that would be reliable in certain benign environments. They foreclose ways of going wrong by producing false beliefs – in particular, they foreclose ways that obtain in a wide set of epistemically possible worlds. Some of these pitfalls may not occur in certain particularly benign possible worlds. This is reflected in our discussion of alternative processes for generalizing from samples. Generalizing from whatever sample comes one's way is reliable only in worlds with particularly simple and homogeneous causal structures. In those worlds, certain ways

of going wrong by incautious generalization simply do not obtain. When one's processes are sensitive to sample size and representativeness, one is more cautious in generalizing – and of course those processes are more robust. The more cautious processes are “conservative” in that they produce fewer new beliefs in response to inputs. So, to insist that it is epistemically desirable for one's processes to be robust as well as reliable in one's world is to require that one's epistemic processes be “relatively cautious” or “relatively conservative.” Again, assuming we have provided for at least a minimal level of robustness, further robustness must be balanced against costs in terms of the productivity of processes – the productive “power” of processes is called for in pursuit of the central epistemic end, along with robustness and reliability.

We have argued that, just as reliability in the agent's world is an epistemically valuable property of processes, so also is the robustness of reliability. The reasons for thinking that robustness is important for the objective appropriateness of processes are closely related to those for thinking that reliability contributes to the objective epistemic appropriateness of a process. Further, when one attends to this dimension of what makes for objective appropriateness of processing, one finds that certain apparent counterexamples to the straightforward reliabilist account can be overcome. While these considerations may establish that robustness and reliability generally serve as coordinate conditions, at least in worlds where each can be had, our formulations remain intentionally noncommittal with respect to details regarding just how these conditions work in tandem. Further discussion must remain for another paper.

#### NOTES

\* We wish to thank Jon Kvanvig and Mark Timmons for helpful comments on earlier versions of this paper.

<sup>1</sup> At least this is the conclusion to which Goldman has been led (1992a). The considerations that have driven most reliabilist thought have always pointed in this direction, so that Goldman's earlier analysis in terms of reliability in “normal” worlds seemed without basis in treating justification as a rigid designator whose reference was to be fixed by what was reliable in a possible world characterized by our subjective understanding of that world. Such an analysis seemed

to conflict with the objectivist thrust of reliablism generally, and to make little sense of the associated externalism. It rendered empty the distinction between our understanding of objective justification and what makes for objective justification.

On the other hand, reliabilists have shied away from a flat-footed position in which reliability of generating processes is itself alone sufficient for objective justification. Our reflections in this paper underscore and elaborate on the reasons for qualifying the reliabilist analysis.

<sup>2</sup> Ultimately, it matters little whether we hold onto the term ‘justification’ here, so long as we are clear about what we are about. Because of that term’s deontological associations, some have felt better in abandoning it for alternatives. On the other hand, there seems a lot to say for holding onto the term – since much work has used the term while being primarily concerned with whatever epistemizes true belief. If readers find the deontological suggestions too difficult to set aside, they are urged to translate our talk of “strong justification” into talk of “warrant” or epistemically objectively appropriate processing.

<sup>3</sup> In the interest of simplicity, we have made little of the place for realizability of processes in their being objectively appropriate for an agent. For some reflection on these matters, see Henderson (1994b) and Henderson and Horgan (forthcoming).

<sup>4</sup> Whatever exactly the systematicity of beliefs involves – it is generally the sort of thing that Kitcher (1989) characterizes in terms of explanatory unification. Such crude pointing will suffice for our purposes here. Generally, the present point reflects the commonplace that, epistemically, we are not interested simply in truth, or simply in the production of isolated true beliefs, but rather in truths that are so related to others as to be interestingly situated in a comprehensive (or at least wide) understanding of the world.

<sup>5</sup> Henderson (1994a) develops this understanding of models of objectively appropriate processing in terms of models of epistemic competence. His development helpfully reflects the objectivist notion being pursued here. However, in being focused in almost exclusively reliabilist terms, it is now seen by its author as incomplete.

<sup>6</sup> It is worth noting, however, that simple reliability has seldom been proposed as alone generally adequate for justification.

<sup>7</sup> This distinction is employed to recognize that thinkers may hold quite different understandings or descriptions and yet be taken to be thinking of (“conceptualizing”) the “same thing.” The point is commonly appreciated in accounts that have learned from so-called causal theorists of language. Because this shared reference looms so large in understanding communication across theoretical differences, and because it plays such a large role in translation, it is common to think of those employing deeply differing descriptive understandings of some thing as (sometimes at least) nevertheless sharing “a concept.”

<sup>8</sup> We realize that the appeal to such scenarios is rendered problematic by recent trends in philosophy of mind, according to which the content of most or all of one’s intentional mental states allegedly depends on certain kinds of actual connections between occurrences of such states in oneself (and/or in one’s evolu-

tionary ancestors) and the actual environment. Those who favor certain versions of content-externalism will doubt whether a brain in a vat has systematically false beliefs. (Some will doubt whether such a brain has beliefs at all, or any other intentional mental states; and some will doubt whether it has any mental states, even qualia.) We lack the space to address this issue in detail here, so we will make just two remarks. First, we think that versions of content-externalism denying that brains in vats have systematically false beliefs are deeply wrongheaded, despite their current popularity; this denial should properly be viewed as a *reductio ad absurdum* of such views. (See the critique of a hypothetical content-externalist called “Strawman” in Lewis 1994, especially pp. 423–25.) Second, we suspect that the points about objective epistemic justification that we will be making in this paper, resting partly on considerations about demon-worlds and brain-in-vat scenarios, probably could also be made by appeal to more complicated hypothetical scenarios that finesse content-externalist considerations. We will not pursue such complications here.

<sup>9</sup> We employ the ponderous formulation, ‘perceptual belief’, in order to steer clear of the success-term usage of ‘perceptions’.

<sup>10</sup> The prospects for a truth-conducive process that generates some class of *a priori* beliefs are as difficult to gauge as is the plausibility of competing understandings of *a priori* truths. Perhaps agents in classical demon-worlds could have truth-conducive processes that generate a very limited set of *a priori* beliefs such as the *cogito* – perhaps.

<sup>11</sup> This source of misleading experience turns on a “theory-ladenness” or “training-ladenness” of (at least some) experience. The actual extent of such theory-ladenness need not be determined for our purposes here – that of characterizing what are epistemically possible worlds. Further, nothing that we need or use here turns on the pessimistic view that the lading of experience with training or theory produces distortions that cannot be overcome. It is worth noting here that the appearances, as what are shared, are not theory-laden in the sense of themselves involving optional, partly socialized, theory. Our talk of appearances reflects our idea that there is “an element” or “dimension” of related, but theory-informed, experiences that is shared and remains in the experience despite experience being colored by theory.

<sup>12</sup> Others have also sought to shed light on judgments regarding reliability by comparing our judgments about reliability in automobiles. For example, Heller (1995) makes use of such comparisons while exploring contextual elements of reliability. Heller’s work, like ours, is rooted in the idea that epistemic judgments that are responsive to epistemic values such as reliability (and, for us, robustness) are not remote from many practical contexts. In such contexts, we make quite nuanced factual and evaluative judgments.

<sup>13</sup> While our gloss on the notion of robustness, in terms of a processing being reliable in a wide set of possible worlds, serves to fix the notion at an intuitive level reasonably well, we acknowledge that this gloss is itself not formally adequate. The limits of the characterization are not simply that it is vague on the score of how wide a set is wide enough for robustness – which does not trouble us

– but also that the notion of epistemically possible world, as developed here, allows for such a fine-grained variation in possible worlds that there may be too many of them. The point was nicely captured in a comment by Jon Kvanvig. Take any intuitively understood possible world – with all the variation that one might think to be epistemically relevant. There would then seem to be infinitely many different worlds like that world – as we can take them to differ in some feature such as the favorite color or favorite number of some individual in that world. It then would seem as if the set of epistemically possible worlds in which a given process is reliable is infinite, if it is reliable in any – and that the set in which it is not reliable is likewise infinite. Presumably there is some formal way in which the characterization might be tightened up, collapsing the worlds differing in epistemically uninteresting detail into single epistemically possible worlds (or world-classes). Obviously, something like this is also needed to make sense of degrees of robustness. We are satisfied that our characterization is sufficient for our purposes in this paper.

<sup>14</sup> This focus may itself be framed in terms of what any epistemic agent *ought* to find valuable features of cognitive processes, given that the agent were to articulately reflect on his or her epistemic ends and projects, and given that the agent also recognized the characteristic epistemic situation of fallibility or uncertainty regarding which epistemically possible world is the agent's.

<sup>15</sup> We should admit that human reasoners can pay little attention to whether their processes are truth-conducive. Many may employ processes unthinkingly, unreflectively, and with little thought to the reliability of those processes. However, insofar as they are epistemic agents, they are concerned with producing or fostering true belief-systems. To will such an end is to will the means to it, as one cannot employ a means without (at least implicitly) believing that it is a reasonably effective means to the relevant end (at least effective relative to the range of means to that end that are open to the agent). Minimally, this requires that epistemic agents would have some tendency, upon challenge or question, to insist that reasoning like that (the sort of reasoning just instanced by them) was a good way of reasoning – and relatively likely to produce true beliefs. The present way of understanding the value of robustness depends only on the claim that cognitive processes are conditioned by such tendencies and understandings in cognizers.

<sup>16</sup> The example of robust and nonrobust processes employed earlier – that of general inductive processes having or lacking sensitivity to possible sample bias – does not illustrate the present point. For, we think that our world is characterized by significant heterogeneity within populations. So, in the world as we take it to be, the feature needed to make for reliability in the face of its heterogeneous populations is also a feature that makes for some robustness. To illustrate the present point, think instead of the way in which scientific work is informed by contemporary scientific understandings of the world. Scientists employ background theory to inform their inferences and their choice of experimental setup. It is often noted that scientists can be enamored of a favored “paradigm,” “research program,” or general theoretical understanding and approach. Scientists may then boldly push ahead, crucially relying on the favored theory in their experimental

design or theoretical reasoning. (As Kitcher 1989 notes, such timidity may look very different when viewed from the social versus individual levels. For simplicity, let us consider the matter at the individual level.) Suppose that Daisy is a scientist who is rather incautiously committed to employing and elaborating her favored theoretical approach. In her reasoning or experimentation, she is relying on the world being one fairly particular way, rather than some range of other ways that are epistemically possible. Predicated as it is on a quite speculative theoretical background, it seems that Daisy's approach will be reliable only (or almost only) if the world is rather like her favored theory represents it as being. Because there are many alternative, epistemically relevant, possible ways that the world might be (not even considering demon-worlds), Daisy's procedure is not robust. It would generally not be reliable in these alternatives. So, even were Daisy lucky – in that the world is like she conceives it in her background theory, so her procedure is reliable in her world – that procedure is not robust.

<sup>17</sup> An analogy: when one's aim is uncertain and when one makes little effort to point one's weapon so as to track one's target, one minimizes one's risk of missing the target by using a sawed-off shotgun rather than a rifle.

<sup>18</sup> Reliability itself has an important counterfactual or modal dimension. The notion of reliability needed by reliabilist epistemology involves a concern for how processes would fare in an environment – where the environment is not understood in terms of just those situations that, as a matter of “accidental” biographical fact that agent happens to get involved in. We might then want to distinguish between local intra-world reliability and global intra-world reliability. The latter is ultimately the concern of reliabilist epistemology. It turns on the truth-related *propensities* of a process within an epistemically relevant possible world. We might then represent the modal dimension of reliability in the agent's world in terms of a set of possible worlds individuated in a more fine-grained way – say in terms of worlds represented as different “trajectories” of an agent within a particular epistemically relevant possible world. (This is nicely reflected in the formulation “non-local or global intra-world reliability.”) All this points to a close kinship between the standard reliabilist concern for (global intra-world) reliability and the concern for robustness (or global safety, or global inter-world reliability of reliability). We will need to pursue this kinship further in a separate paper.

<sup>19</sup> We should express a doubt regarding the success of the pragmatic justification of induction on the score of whether it really managed to show what it sought. To make the case, it was simply supposed that induction would catch onto any regularity. This supposes an amazingly powerful and sensitive cognitive system engaging in the induction. It abstracts away from much that would make for any concrete implementable cognitive process – and thus much that probably limits what could be claimed for induction as a cognitive process. These misgivings parallel those expressed by Cherniak regarding the treatment of standard formal logic as straightforwardly providing epistemic standards. However, for reasons that harken to Goodman, the sort of abstraction encountered in the pragmatic justification of induction is, if anything, more problematic.

<sup>20</sup> A more nuanced analysis of the pragmatic justification might proceed in terms

an indexed notion of robustness – that of the “robustness of methods for certain classes of epistemic tasks.” Thus, consider the processes that will allow us to generate a particular sort of belief – say generalizations, to employ the crude typology of the present example. We might imagine various processes for doing this, some of which work – that is, generate true components of a belief-system – within no world; others work within a small set of possible worlds to which they are tailored; and yet others might work within a wider set of worlds. The latter will be robust-for-a-process-of-that-sort. Compare this to the idea of a baseball player who “runs well for a catcher.” Or of a racing car that is dependable for such cars. A catcher with this property is *ceteris paribus* to be valued, even though most teammates would have more success in beating out bunts. A car with the above property might be too temperamental to serve as everyday transportation.

<sup>21</sup> Two qualifications seem called for. First, because robustness comes in degrees, one employing a robust process may not have strictly minimized risk, but would have done something like satisficing with regard to acceptable risks. Second, ultimately, the agent will have satisficingly-minimized risk without getting unduly or unliveably conservative and agnostic. It is well known that there are always trade-offs between minimizing risk and employing productive processes. We have written of the epistemic end of fostering true belief-*systems*. Reliability and robustness speak most directly to the production of true beliefs. However, we also want a comprehensive system of beliefs. These concerns commonly pull us in differing directions and must be balanced.

<sup>22</sup> Again, this requires qualification. Should we say that had they not landed in a demon-world, their processes would likely have been reliable? The basic idea seems right, but the relevant notion of probability may be difficult to make out with precision.

<sup>23</sup> Of course, thinking about evil-demon-worlds has led others who are sympathetic to externalism and reliabilism to refine the basic reliabilist approach. Sosa’s (1991) response to what he terms “the new evil-demon problem” provides one prominent example. However, our response differs significantly from Sosa’s. Sosa simply relativizes judgments of “aptness” to worlds – so that, in our earlier illustration, Constance’s thought would be apt relative to our world, but not relative to demon worlds (while Faith’s thought would not be apt relative to either). We believe that BonJour is correct in insisting that this really is not an adequate response:

For surely the main intuition is that the demon victim’s beliefs are justified without qualification in the environment that he inhabits, not merely that they are justified in relation to a quite different environment whose relevance to his actual epistemic environment is pretty obscure (1995, p. 211).

Notably, our response is to identify a non-world-relative objective feature of processes, a feature that can contribute to the epistemic value of processes without qualification in the agent’s actual environment. We are thus able to honor fully the intuitions to which BonJour appeals, and to do so by appeal to a fully objective feature of cognitive processes.

<sup>24</sup> Insofar as appeals by reliabilists to an intra-world meta-reliability requirement are motivated by an aversion to an epistemic role for dumb luck, robustness of belief-forming processes is a condition that addresses this motivation better than meta-reliability itself does. Swamp Thing, who is just like an ordinary human except that he came into being as a result of random spontaneous chemical interactions when a lightning bolt struck a swamp rich in organic molecules, could well have epistemically impeccable belief-forming processes, even though these processes were generated by utterly unreliable *meta*-processes. Swamp Thing's belief-forming processes will be impeccable provided that they are *robust* as well as reliable. (We realize that the appeals to Swamp-Thing Scenarios are rendered problematic by certain prominent recent trends in philosophy of mind, according to which a creature who lacks a suitable evolutionary pedigree would thereby lack intentional mental states (or would lack mental states altogether, even qualia). But our remarks in note 8 about brain-in-vat scenarios apply here also, *mutatis mutandis*.

<sup>25</sup> A further thought-experiment seems to reinforce the conclusion that the robustness condition applies to cases even when there are reliable processes available. There are epistemically relevant possible worlds in which either reliability or robustness can be had (at least in principle), but where no process would be both reliable and robust. This would obtain in some non-classical demon-worlds, those in which there is a rigid demon. Because the rigid demon provides the agent with *systematically* misleading appearances, there may be a special-purpose, systematic correction process possible, at least in principle. Envisioning such a process is a matter of envisioning the demon-world (in particular, its structure of systematic deception), then envisioning a belief-forming process that treats the appearances as systematically deceptive in exactly the way they happen to *be* deceptive, and systematically corrects for these appearance in exactly the right way. To come to implement such a process, an agent would need to adopt it in a fashion that, in the nature of the case, must be counter to appearances, arbitrary, unmotivated and unmotivatable. The odds of an agent just "hitting on" the reliable strategy in such a perverse fashion are vanishingly small. Suppose, however, that an agent in a rigid-demon world *does* hit on exactly the right strategy – blindly and miraculously, purely as a matter of epistemic dumb luck. Such an agent could employ these *de facto* reliable belief-forming processes, or instead could employ robust processes (which happen not to be reliable, although the appearances provide no clue of this), but could not do both at once. Surely the *epistemically appropriate* processes, for such an agent in such a rigid-demon world, would be the robust ones rather than the reliable ones.

For what it is worth, we are inclined to see judgments about such extreme and fanciful thought-experiments as rather less weighty than the general considerations that point to the uniform applicability of robustness. It is also worth nothing that robustness is particularly uniform in its application to agents in various possible worlds because – unlike reliability – it does not vary with possible worlds. A process is robust (or it is not) no matter what world the agent happens to be in. Thus, while there are possible worlds in which there are no reliable

processes available to agents, and reliability seems to there drop out, robustness is not dependent on the particular world that the agents are in.

## REFERENCES

- BonJour, L. (1995): 'Sosa on Knowledge, Justification, and Aptness', *Philosophical Studies* 78, 207–220.
- Chernaik, C. (1986): *Minimal Rationality*, MIT Press.
- Davis, Lawrence H. (1974): 'Disembodied Brains', *Australasian Journal of Philosophy* 52, 121–132.
- Evans-Pritchard, E. (1937): *Witchcraft, Oracles and Magic Among the Azande*, Oxford University Press.
- Geertz, C. (1983): "'From the Native's Point of View": On the Nature of Anthropological Understanding', in G. Geertz (ed.), *Local Knowledge*, Basic Books.
- Goldman, A. (1986): *Epistemology and Cognition*, Harvard University Press.
- Goldman, A. (1992a): 'Strong and Weak Justification', in A. Goldman (ed.), *Liaisons*, MIT Press.
- Goldman, A. (1992b): 'Epistemic Folkways and Scientific Epistemology', in A. Goldman (ed.), *Liaisons*, MIT Press.
- Heller, M. (1995): 'The Simple Solution to the Problem of Generality', *Nous* 29, 501–515.
- Henderson, D. (1994a): 'Epistemic Competence', *Philosophical Papers* 23, 139–167.
- Henderson, D. (1994b): 'Epistemic Competence, and Contextualist Epistemology', *Journal of Philosophy* 91, 627–649.
- Henderson, D. and Horgan, T. (2000): 'Iceberg Epistemology', *Philosophy and Phenomenological Research*, 497–535.
- Kitcher, P. (1989): *The Advancement of Science*, Oxford University Press.
- Latham, Noa (forthcoming): 'Chalmers on the Addition of Consciousness to the Physical World', *Philosophical Studies*.
- Lewis, D. (1994): 'Lewis, David: Reduction of Mind', in S. Guttenplan (ed.), *A Companion to the Philosophy of Mind*, Blackwell.
- Nagel, Thomas (1974): 'What Is It Like to Be a Bat?', *Philosophical Review* LXXXIII, 435–450.
- Sosa, E. (1991): *Knowledge in Perspective*, Cambridge University Press.

*Department of Philosophy*  
*The University of Memphis*  
*Memphis, TN 38152*  
 USA