Explanation and Rationality Naturalized[1]

David Henderson

Draft 4/26/09 1:58 PM, Not for Quotation

I. Issues

When explaining some action or thought of an agent, one often notes antecedent beliefs and desires that would make that action or thought rational. Some have suggested that finding agents rational provides the "only clear pattern of explanation that applies to the mental" (Davidson, 1982, p. 299), and that this model is equally central within the human sciences generally—making for a explanatory practice that can neither be significantly refined nor transcended without loss of the subject matter of those sciences (Rosenberg, 1985, 1988).[2] I have elsewhere argued that is it not findings of rationality per se that are explanatory; rather, what is explanatory are findings of ways of reasoning (cognitive processes) that conform to what we have come to know and expect about human cognition (Henderson 1993, 2002, 2007).[3] Of course some forms of rationality may be expected, and recognition of instances of these may then be explanatory. But, the same can be said for findings of certain forms of irrationality. (Of course, the forms of rationality and irrationality that we expect in particular cases is conditioned by the training we understand an agent or agents to have undergone, and by salient aspects of the circumstances.) This said, there remains a matter that has not been adequately investigated: whether or not not rationalizing explanation might be better understood when thinking of rationality in a naturalized fashion, rather than along more traditional lines. Most writings on rationality and explanation have supposed that what makes for rationality (and explanation) is determinable a priori. More guardedly put, in discussions of rationalizing explanation the understanding of what makes for rationality has been drawn from disciplines that can and do proceed independently of significant empirical information regarding humans as a class of limited cognitive systems. Commonly, the understanding of rationality has come from some combination

or first-order logic, the probability calculus (including Bayes' Theorem), and decision theory. These have been thought to provide some reasonable measure for what would be objectively rational. However, in epistemology, and increasingly in accounts of normative rationality more broadly, there has been the realization that fitting normative standards for epistemic and practical rationality need take into account the resources of real cognizers—typically humans as a class of cognitive systems (Quine, 1969; Harmon, 1986; Cherniak 1986; Goldman 1976,1986, 1992a, 1992b, 1999; Bishop and Trout 2005, Henderson and Horgan 2008). Does the new realism in accounts of rationality (associated with naturalized epistemology) itself significantly improve the prospects for unproblematic forms of rationalizing explanation? Do earlier misgivings about rationalizing explanation ring hallow when the rationality to be attributed is naturalized? I here argue that, while explanation in terms of naturalized rationality would be free of one fatal flaw had by explanation in terms of rationality understood in the traditional fashion, it would yet have parallel flaws.

II. Explanation, some widely shared points:

I want to begin with some common ground shared by alternative accounts of explanation. The misgivings about rationalizing explanation developed in this paper rely only on this common ground, and thus do not presuppose any one approach to explanation.

Consider a true causal statement that also explains some event. It might be explanatory because it connects with background nomic generalizations or invariant generalizations (as in subsumptive accounts of explanation), or because it points to a causal relation or process (as in ontic accounts), or because it answers or allows answers to a how-question, why-question, or what-if-things-had-been-different-question (as in erotetic approaches to explanation).[4] But, by all accounts, the *explanatory character* of the claim is dependent on something beyond that the *truth* of the claim. Causal statements are referentially transparent—their *truth* is preserved under

substitutions of coreferential terms. In contrast their *explanatory character* may be said to be referentially opaque. Explanation is sensitive to how events are characterized. Explanations call attention to patterns of actual and counterfactual dependency. Some characterizations of pertinent events, objects, or systems call attention to, or make evident (perhaps in light of background information) patterns of counterfactual dependency. Others do not. Thus, if one substitutes within an explanatory singular causal claim in a way that does not change the events related there, the product will be another true causal claim—but it need not be an explanatory claim. Suppose that my release of blood into the waters causes there to appear on the scene a bunch of sharks. Blood is red, and red is my daughter's favorite color. So, the event that is my release of blood at *t* is my release of red liquid at *t*, and is my release of a liquid of my daughter's favorite color at *t*. So, if my release of blood at *t* caused the congregation of sharks, my release of a red liquid at *t*, and my release of a liquid of my daughter's favorite color at *t*, caused that congregation. The three causal claims are each true, provided the first was, and provided the singular causal claims are understood in the standard event-relating ways. But, only the first is explanatory. Only it gets at the pattern of dependencies. Sharks (we may suppose) are differentially attracted to traces of blood in the water, but not to the color of liquids in the water, or to features high in my daughter's preference orders. This differential response makes for the counterfactual dependencies obtaining in cases—dependencies on which successful explanations turn. This much is crucial to the explanation of events.

It is worth noting that this point—that explanations call attention to patterns of counterfactual dependency—is not restricted to causal explanations (although those explanations that concern us here are causal explanations). For example, in functional explanations—explanations in which one explains how a system manages some task by the organized operation of simpler processes—one is given to understand that (a) that were one of these simpler processes not to perform their input-output chores on cue, the system would fail to work as it commonly

does, and (b) were the simpler systems to have different input, their output, and thus the output of the larger system, would have been different in specific ways (Cummins, 1983).

It should not be controversial that reason-citing explanations are a form of causal explanation. But, my reasoning here will not turn on philosophical points regarding causation or causal explanations. I hope to make my arguments relying only on the common coin: that explanations of what happened in cases turn on revealing patterns of actual and counterfactual dependencies. Explanations in terms of reasons are not explanatory monstrosities—but rather, common citizens of the realm of explanations. Explanations in terms of reasons work by calling attention to features of the agent's (or the agents') antecedent beliefs and desires on which the agent's (or agents') doing or thinking is counterfactually dependent.[5] Such explanation turns on patterns of counterfactual dependencies. This is common ground. The crucial questions thus are: what is the character of the counterfactual dependencies obtaining between the agent's beliefs and desires and the explanandum particular thought or action?

IV. A venerable target:

Many have imagined that belief-desire explanations work by exhibiting the explanadum thought or action as rational in view of antecedent beliefs and desires of the agent or agents. Supposedly, upon appreciating that those antecedent thoughts would have provided a basis on which the agent might have rationally arrived at the subsequent thought or decision, one infers that the agent arrived at that result in those ways. On the other hand, mention of antecedent beliefs and desires that would not make the thought or action rational may leave one wondering just what moved the agent to think or act in the way in question. Or so the story goes. Supposedly, we are stuck with this form of explanation—until we give up on reasons explanations altogether.

There is a difference between being rational and merely seeming subjectively to be rational. To explain by exhibiting rationality thus requires that such explanation be informed by principles of rationality. More fully: the rational principles or standards that inform such explanation are (1) *normative principles*—not empirical results about human approximations to how one objectively ought to reason, and (2) *principles of objective rationality*—not merely norms embraced by this or that individual or group, not merely subjective rationality. That the principles of rationality supposed here are to characterize objective rather than local norms is reinforced by the associated idea that these same norms constrain interpretation, and thus will be found to largely honored universally (among those with beliefs and desires) (again, see Davidson 1980a, 1980b; Rosenberg 1988, chapter 2). I take (1) and (2) to generically characterize rationalizing explanation, properly so called. Our concern in this section and that to follow is with what might be termed the received account of rationalizing explanation.

On the received approach, a priori reflection is looked to as the source of the objective normative principles. This is a somewhat flatfooted way of putting the point—for some proponents of the idea in its received form have not been comfortable with the idea of a priori knowledge or justification. Nevertheless, they have thought of the relevant normative principles as being drawn from a set of inquiries that have commonly been candidates for a priori sources—kinds of inquiry that are pursued independent of psychology and cognitive science. Central here have been (a) first-order logic and associated demands for consistency and respect for implications, (b) the probability calculus, including Bayes Theorem, and (c) decision theory, commonly calling for something like utility maximization. For ready labels, let us call such empirically uninformed normative models of rationality, *apriorist principles,* and let us call the standard or received conception of rationalizing explanation *the apriorist approach.*

Even when thinking in terms of apriorist principles, the terms ‘rational’ and ‘rationality’ are used in a variety of ways. Thinking just of epistemological rationality for the moment, one talks

of rational beliefs, or agents being rational in believing, and of rational processes. Crudely, and supposing the apriorist approach,[6] we can highlight the following common ways of talking. For a *belief* to be rational in view of certain antecedent states of an agent is for the content of those antecedent states to provide sufficient support for the belief. This is a matter of relations between contents. (For the apriorist, these relations are those characterized by logic, mathematics, and related fields.) Thus, to say simply that a belief is rational for an agent to hold is not to say anything about the processes by which the agent came to hold that belief. To say that an *agent* is rational in holding a belief can be understood in two ways. An agent is *weakly* rational in holding a belief just in case a belief with that content is rational, given the agent's antecedent states. An agent is *strongly* rational in holding a belief if the agent came to hold that belief in a way that depended on its being rational, given the agent's antecedent states.[7] This is a matter of the cognitive processes in play in the generation of the belief being processes that are sensitive to the demands of a priori rationality. *Processes* are said to be rational when they are sensitive to and respect the contentful relations making beliefs rational.

The idea that one can explain a thought or action by exhibiting its (apriorist) rationality is most charitably understood as the idea that one explains by exhibiting the agent or agents as strongly rational in that thought or action. More will be said below, but anticipating a little: the idea that one explains by exhibiting mere weak agent rationality would seem to be completely without promise. Because weak rationality does not involve any constraint on the processes in play in the episode, and because patterns of dependency on which explanation turns are a function of the processes in play in the relevant episode, exhibiting mere weak rationality would seem not explanatory. (Admittedly, the exhibition of weak rationality might plausibly play an evidential role in an abductive argument with the conclusion that, probably, the agent is there strongly rational.)

V. Critique of the apriorist model:

It is, or should be, clear that humans do not fully consistently conform to aprioristic normative models. Moreover, they fail to conform in such systematic ways as to make clear that these models do not characterize pervasive cognitive tendencies. In thinking about the ways in which humans fail to conform to the demands of objective rationality, it is important to distinguish rationality and rational competence, competence and performance, competent performance and performance errors. It is also helpful to consider rational competence in rudimentary forms and those forms that come by training. In what follows, I explore how such distinctions bear upon the idea that one can explain an action or thought by showing that it was rational in view of antecedent cognitive states of the agent or agents. The present section focuses on the distinction between rationality (weak and strong) and rational competence (and rationally competent performance). When this distinction is drawn by an apriorist, it looks rather different from what a more naturalized thinker would envision—and the difference is crucial to the argument of this section.

A competence—for example, a grammatical competence, or a rational competence, or a perceptual competence—is a complex capacity. A competence is constituted by a select set of dispositions or processes that can be put in play within a system of a kind. Typically, normal adult humans have provided the relevant kind—thus we study human linguistic competence or human reasoning competence. A competence is a descriptively constrained normative ideal that informs thinking about correct performance—talking well, reasoning well—and accounts for how humans manage to systematically produce correct performances in a reasonably wide range of cases. A competence is constituted by dispositions to those tractable processes by which a class of cognitive systems can manage well (or optimally) their cognitive (or other) chores. There are two things to emphasize here: the tractability of processes (for the class of systems in question) and their contribution to managing well the tasks or chores in question.

Focus first on tractability. For our purposes, we can think of processes, actual and possible, as characterized abstractly in terms of inputs and out-puts, in terms of causal roles. Humans can employ some processes and not others. For example, humans typically can employ a range of processes for object location that turn on making use of visual, tactile, and some auditory input. Eco-location is not tractable for humans, while it is tractable for bats. Bat competence for locating objects includes dispositions to processes of eco-location. Human competence does not. Normative epistemic models for how typically endowed humans ought to engage with their environment include various processes using visual, tactile, and auditory inputs—these constitute a component of human epistemic competence. A human epistemic agent would be remiss to neglect such information and processing. On the other hand, one does not think that a human agent is remiss for not employing eco-locating processes. A normal bat that omitted to eco-locate would not have fittingly cognitively engaged with its environment. On other cognitive chores, humans find tractable certain processes that are beyond bats. Humans can be sensitive to sample characteristics when generalizing in ways that bats cannot. Processes that then modulate generalization from samples may be part of human epistemic competence—but it (and articulate generalization from samples) is not a part of bat competence. The dispositions to processes constituting a competence may be innate or acquired (perhaps after significant training). In either case, competence is a matter of a select set of (dispositions to) tractable processes.

Competence is constituted by processes by which one can manage well the relevant tasks or chores. Focus now on human competences—such as linguistic competence, competence in riding a bike, in fielding in baseball, perceptual competence, epistemic competence more generally, or rational competence (our concern here). While some abstractly possible processes are not included in the relevant select set because they simply cannot be realized in humans, other processes, even actual processes in humans, are not included in the relevant select set because they are irrelevant to the endeavor of concern—for example, the process of mathematical

induction is probably not a part of either baseball fielding competence or bike riding competence. Other actual processes are not a part of competence because they interfere with the operation of the select set and with associated success in the endeavor of concern: the process of falling asleep is not a part of driving competence.

Dispositions to the select processes amounts to a capacity. Putting these processes to work, using the capacity, undergoing the processes, accounts for normatively correct performance—where it is found. (The remainder of the present section can be understood as developing the implications of this last sentence.)

In part because a competence is constituted of a select set of dispositions or processes, an agent may be competent and yet may have other or additional processes in play in a given case. These may interfere with the workings of competence. (For example, highly stressful conditions may evoke processes that interfere with the agent's linguistic, rational, or even bicycling, competence. Thus, competent agents can make performance errors.

One issue with far reaching implications is whether, or to what extent, what makes for *correct performance* can be understood independent of, or prior to, understanding what amounts to *competent performance*. Rationality, strong rationality, is correct performance in the production and maintenance of beliefs, desires, or decisions. Whether what is rationally correct can be understood prior to understanding what makes for rational competence is an issue on which apriorists and naturalists differ. Subject to some qualifications, the apriorist would hold that an understanding of rationality or rational correctness can be had prior an account of human epistemic competence. The naturalist denies this. Let us consider how the matter—the relation between correct performance and competent performance—looks in other domains.

For some matters, the competence in question may largely constitute what is correct performance. Plausibly this obtains in the case of linguistic competence. Competence with (some

dialect of) English might itself characterize what is correct linguistic performance using that language or dialect. There is, it seems, little daylight between correct performance in some dialect and competent performance.[8]

It is certainly true that for many forms of competence, the members of select set of processes that constitute the competence qualify by virtue of being (a) tractable for the cognitive system, and (b) conducive to some characteristic end state. The characteristic end state provides a *telos* that informs the evaluation of candidate processes and performances. In such contexts, there is at least a notion of "good" that is prior to the notion of competence. But, to recognize that there may be a notion of good that informs the evaluation of processes (selecting those that constitute a kind of competence) and of performances (as correct) does not imply that what makes for correct performance can be independent of the relevant competence. Consider a sports example: competent fielding in baseball is a matter of employing processes that tend to achieve certain ends (catching balls in flight and those skidding along the grounds, distributing the ball to others in a way that makes probable certain results). Here, the ends inform the selection of a set of processes we want in a baseball fielder—fielding competence. What of correct performance? Clearly, satisfying the *telos* in a given cases, is not sufficient for correct performance in that case. Suppose that a ball is hit in the air in the general direction of some child. Suppose that the child reacts by closing his or her eyes and raising his or her glove (perhaps to cover sensitive areas). Finally imagine that by shear luck the glove snags the ball (for an out). *Telos* achieved. But the child did not perform well. Further, achieving that *telos* in a given case is no more necessary for having performed correctly than it is sufficient for correct performance in the case. The highly competent fielder may do all the correct things and not "make the play"—notably, when the play could not be humanly made. What would make for a correct performance? The answer seems to be a kind of competent performance in which the agent employed a select set of processes that conduce to achieving the *telos*. So, in this sports context, while there may be a notion of "good"

that is arguably prior to the that of both correct performance and competent performance, these latter notions seem intimately connected.

Something along these lines might hold for the evaluation of belief-forming processes as components of an epistemic/rational competence, and of epistemic performances as cases of epistemic rationality. Certainly naturalized epistemologists are committed to something like this. The naturalized epistemologist would insist that correct performance is competent performance—both being a matter of using processes that are appropriate to humans (or the class of cognitive systems being evaluated). We will return to the naturalized epistemologist shortly. First, focus on the apriorist. On the apriorist model, what would constitute correct cognitive performance (specifically, epistemic rationality) can diverge significantly from competence (specifically human epistemic competence).

Epistemic competence is relative to a class of cognitive systems. (Not surprisingly, the only class of cognitive systems that is much discussed is humans, typically normal adult humans.) Human epistemic competence is a select set of those processes that can be implemented in humans. It cannot involve processes that humans, even with the best of training and motivation, cannot reliably manage. But, clearly, many traditional understandings of epistemic or practical rationality have envisioned abstractly characterized possible processes (classes of transitions or inferences) that humans could not consistently manage with the best of training and motivation. Such prescriptions are likely when one supposes that one can ascertain by reflection (without drawing on empirical information) what makes for correct epistemic performance. For example, evidentialists would say that there are a priori discernable support relations that may obtain between contentful states. For the evidentialist, epistemic correctness is a matter of processes that would conform belief to the evidence—whatever transitions or inference it would take to systematically manage to do that, those are demanded. Inferential transitions of sorts that would provide logical consistency, probabilistic consistency, the full use of information, and presumably

some nuanced form of abduction—are all demanded. The traditional apriorist understanding of rationalizing explanation has tended to assume such a conception of rationality. Here, sensitive, systematic, ongoing conformity to the possessed evidence (support relations between contents) is taken to provide an a priori standard for epistemic correctness—one independent of the processes making up human epistemic competence. Further, it is highly plausible that humans are not capable of processes that fully and systematically manage this.[9] When such an apriorist understanding of rationality calls for such inferential processes, while competence can involve only (even the best possible constellation of) tractable processes, apriorist rationality and epistemic competence must diverge. The competent agent, using the best select tractable processes, and employing them without interference or compromise, may thus fail to do what is aprioristically rationally correct—because these processes may not be of sufficient power and discernment to perfectly conform one's belief to the evidence.

To generalize: those who think that epistemic correctness or (broader) rationality can be determined a priori (or at least independently of empirical information about human cognitive capacities and processes—by logic, the probability calculus, and utility maximization, for example) are committed to the view that epistemic or rational competence can (in honesty, will surely) fall systematically short of what is rational. For the apriorist, the strongly rational agent would need to be *super*-competent. On this approach to objective rationality or cognitively correct performance, there will be "plenty of daylight," plenty of divergence, between what would make for strong rationality (a priori determined) and rational competence (a matter of what humans can manage with the most conducive training and highest motivation).

Some will suspect that the present discussion is saddling the apriorist rationalist with an implausible commitment in saying that apriorist (strong) rationality requires a superhuman competence. Surely, for the apriorist rationalist, one can be rational without being super-human (or without being infinitely cognitively powerful). The apriorist can find

flexible ways of thinking of what is rational. Doubtless, an apriorist rationalist can distinguish a sense in which agents might be rational that is intermediate between being weakly and strongly rational. In a given episode an agent might manage to be weakly rational by virtue of a human rational competence—call this being *competently rational*. Being competently rational is intermediate between the other forms—it requires more than merely being weakly rational, but does not require super-human competence. What makes for competence rationality—even for an apriorist—cannot be understood solely on the basis of those forms of inquiry that have been the stock sources for the apriorist. The idea that one might explain an agent's thought or action by exhibiting the case as one in which the agent is competence-rational will look increasingly attractive over the course of the present section—it will then be the focus of the following section.

Notice that, for the apriorist, competence rationality requires that the agent be weakly aprioristically rational in the case (or in the limited range of cases in question)—and that this weak rationality result from the operation of cognitive processes that are appropriate to humans, from human rational competence. It is worth also noticing that, while it might be imagined that competence-constituting processes are just some subset of strong-rationality constituting processes (perhaps with provision for limited memory and time) this is not necessarily the case. For example, in the empirical literature, one finds the suggestion that certain heuristics, or certain "fast and frugal processes" might be a part of competence.[10] It is not as though such processes would amount to simple and undemanding ways of getting to exactly what weak rationality would require across the cases on which they might be put to work. For significant ranges of applications in cases, the results to which these heuristics would give rise may be different from those that the apriorist would take to be strongly or weakly rational. So, not all the results of rational competence would be cases of being "competently rational" as characterized above. This is because being

competently rational was characterized in a way that required weak apriorist rationality. The point is that the apriorist and the naturalist thing about competence rationality in importantly different ways. A agent may be said to be *competently aprioristically rational* when that agent's thought or decision (a) is the result of rational competence, and (b) happens in that case to be weakly aprioristically rational. In constrast an agent will count as being *competently naturalistically rationality* when employing (without significant interference) human rational competence. For a naturalized epistemologist, if ideal human rational competence takes one there, it is rational to go there.

Naturalized epistemology (and a naturalized approach to rationality generally) contrasts with the apriorist approach in a way that will be highly significant for the argument of this section (As a result, the fatal flaw in the apriorist approach to rationalizing explanation here diagnosed will not be shared by an approach in terms of naturalized rationality.) Naturalized epistemologists insist that rationality, or rational correctness in general, is not to be understood as independent of what humans (or the relevant class of cognitive systems) can manage.[11] One cannot determine by a priori reflection what are the requirements of rationality. Without such empirical information, one cannot determine with precision what cognitive processes make for the epistemically fitting and justificatory ways of forming belief. One might reflectively identify certain general parameters making for objective epistemic rationality. Plausibly some form of cognitive tractability and some form of reliability might be found to be conceptually central to what makes for epistemically objectively rational ways of forming belief. (Reliabilist epistemology is representative of naturalistic epistemology.) But this does not give any concrete account of how one ought to form beliefs—what specific ways of reasoning constitute objectively epistemically rational ways of forming beliefs. With regard to how humans ought to reason—ought to form beliefs, revise preferences, decide to act—the naturalized epistemologist insists that "ought" implies "can." To say that one ought to reason in a certain way implies that humans could, *at*

least with associated training and very high motivation, manage to so reason. Thus, one cannot determine just how humans ought to reason without facing empirical issues regarding what kinds of processes are tractable for humans. On this approach, then, what makes for objective epistemic rationality cannot be understood independent of, or prior to, an account of epistemic competence. There is then little or no daylight to be found between epistemic competence and epistemic rationality.

Of course, the naturalist can draw a distinction between strong and weak rationality. *Strong naturalistic rationality* is a matter of getting to one's thought or decision by way of competence-constituting processes. It is a matter of naturalized competence rationality. *Weak naturalistic rationality* is a matter of getting to a thought or decision that could be gotten from antecedent states by way of competence-constituting processes—although it does not require having gotten to those results using those processes. There is no difference between strong naturalistic rationality and naturalistic competence rationality.

With all this stage-setting, the challenge to the apriorist can be formulated as follow:

1. (A very general point concerning processes and what is explanatory.) Explanation has to do with patterns of actual and counterfactual dependency obtaining in the episode in question, with answers to what-if-thing-had-been-different-questions, and these are a function of the processes there in play.
2. One cannot explain by exhibiting the belief in question to be strongly apriorist rational—for that requires that it be the result of processes that no human can implement. No human is ever strongly apriorist rational. The patterns of counterfactual dependency, the answers to what-if-thing-had-been-different-questions, will always reflect processes of a more limited sort.

3. At best the processes at play will be those making for human rational competence. But, to explain in terms of competence rationality would be to abandon the apriorist approach and move towards a more naturalized approach—and the pattern of dependencies in the case would not mirror or track aprioristically rational relations.
4. Weak rationality is not explanatory—because it does not entail anything about the processes in play. Weak rationality does not entail patterns of counterfactual dependency, answers to what-if-things-had-been-different-questions, or causal relevance.

These points leave us with the possibility that rational competence, or some form of competence rationality, might be explanatorily relevant—an issue to be addressed in the following section.

What can be explanatory is a matter of the actual processes in play in an episode. Only these processes—the transitions to which they do and would give rise, the input on which they work in an episode, the features of that input to which those processes are sensitive—can be explanatory of what transpired in a given episode. Patterns of counterfactual dependency, relevance to what-if-things-had-been-different-questions, causal relevance, all depend on the actual processes at work in an episode. Mentioning abstractly possible processes that are not possible in the kinds of system with which one is concerned certainly cannot be explanatory. Neither can possible processes that such a system could manage, but did not have in play. Thus, what explains what beliefs and desires are formed, or what decisions or choices are made, are processes there at play in agents (or input to those processes to which those processes are sensitive, or aspects of that input to which such processes differentially respond). At best, these processes are among those constituting epistemic competence. So, in the best of cases, explanation will turn on rational competence (the capacity) and on the agent being competently

rational (the agent's performance being controlled by that competence)—not on strong apriorist rationality.

An agent may be weakly aprioristically rational, or not. An agent's cognitive performance might be competent, or not. There are then four possibilities. They are illustrated in Table 1, with examples that suppose common apriorist rationality.[12]

	Weakly Aprioristically Rational	Not Weakly Aprioristically Rational
Result gotten by competent processes	1. The agent draws an obvious conclusion—by employing simple logical techniques that are a part of limited human competence.	3. Using a heuristic, and using it in a manner that is not likely to yield systematically distorted results, the agent arrives at a result that is different from what would follow from very difficult calculations drawing on one's background information.[13]
Result gotten by other processes	2. By wishful thinking the agent arrives at the belief that a certain feared combination of conditions is unlikely—where this result that could have been gotten by multiplying the antecedently known independent probabilities of those conditions.	4. Noticing that a six has not turned up in the last 10 tosses of a die yet thought to be fair, the agent concludes that the probability of a six on the next toss is significantly greater than 1/6—perhaps even approaching ½.

Table 1

Since weak rationality is clearly necessary for rationalizing explanation, we can consider just cases of types 1 and 2.

Here is the crucial point regarding type 1 cases. Correct answers to what-if-thing-had-been-different questions, the truth about patterns of factual and counterfactual dependencies, causal dependencies, are a function of the actual processes in play in the agent or agents in the episode in question. If the processes in play are components of competence, *and if, as the apriorist must hold, competence falls short of rationality,* then the counterfactual dependencies between states will reflect competent transitions rather than aprioristically rational transitions. The answers to what-if-things-had-been-different-questions will reflect the fact that the agent's or agents'

conclusions or decisions were, and would generally be, naturalistically competently rational, the outcome of rational competence, sensitive to whatever rational competence would be sensitive to, but not generally aprioristically rational (weakly or strongly). In effect, in cases of type 1, the agent's or agents' rational competence screens-off apriorist rationality from explanatory relevance—from relevance to patterns of counterfactual dependency. In these cases, where the agent is rationally competent, apriorist rationality is not explanatorily relevant.

Cases of sort 2 are even more straightforward—here the idea of explaining by showing that the explanandum accorded with apriorist rationality (given the agent's antecedent states) has no plausibility to begin with. Of course, the processes actually in play must be tractable processes. These would be neither components of competence (by stipulation of the case) nor processes that make for objective rationality of any stripe. When such processes are in play, the counterfactual dependencies between states will generally reflect neither competent transitions nor aprioristically rational transitions. The answers to what-if-things-had-been-different-questions will reflect the fact that the agent's or agents' conclusions or decisions were (and would be) both rationally incompetent and generally aprioristically irrational. That the belief then happened to accord with what would be aprioristically rational, that the belief happened to be weakly rational, is clearly not explanatory.

These results provide a decisive objection to the apriorist approach to rationalizing explanation. There seem three take away lessons (in addition to the repudiation of aprioristic rationalizing explanation): First, it will never be sufficient for explanation simply to show that that the explanandum is weakly rational—regardless of whether one thinks in terms of apriorist rationality in naturalistic rationality. Second, if the agent's being objectively rational is to have any explanatory role, it must be some form of competence rationality—it must be because the agent was rationally competent and was instancing that competence in the episode in question. Third, the fatal flaw concerning apriorist rationalizing explanation discussed here turned on the

"daylight" between apriorist rationality and human-competence rationality. So it would seem that they might not have force against one adopting a more naturalized approach to rationality. The hunch that the naturalized approach to rationalizing explanation might have fewer problems than the apriorist approach is now partially vindicated.

The central flaw in the apriorist understanding of rationalizing explanation had to do with cases of type 1: cases in which the agent was competent and the result was weakly aprioristically rational. Recall the event that was both my release of blood and my release of a red liquid. The first characterization of the event is explanatorily relevant. The second seems screened off from explanatory relevance. At issue in this section was whether the aprioristic rationality of the result was explanatorily relevant or whether the agent's competence was explanatorily relevant, or both—given that the agent's drawing of a conclusion/decision may be characterized both ways. Patterns of counterfactual dependency, given the sorts of processes in play, were taken to be decisive. The patterns of counterfactual dependency will reflect the processes in play—competent transitions, the aspects of antecedent states to which competent processes are sensitive. So explanatory relevance will lay with the agent's competence, not with aprioristic rationality. Aprioristic rationality explains thought and action about as well as the addition of liquid of my daughter's favorite color explains the ensuing congregation of sharks.

In the background of the above discussion has been an intriguing contrast: for one employing the naturalized approach, ideal human rationality does not diverge from naturalistic competence rationality. A human agent is strongly naturalistically rational in a given thought or decision if and only if that agent is ideally competently rational. This suggests a plausible direction in which one might look when attempting to make sense of the idea that one can explain by exhibiting rationality: on a naturalized understanding of rationality, in a case involving a competent agent exercising that competence without interference, the patterns of counterfactual dependency would seem to support the suggestion that both rational competence and naturalistic rationality are

explanatorily relevant. At least the flaw just charged to the apriorist does not apply to the naturalist.

While we have considered the implications of the distinction between rationality and rational competence, we have yet to factor in variations in agents' *degrees* of competence that can arise from differences in suitable training. Focusing on this last matter yields a misgiving regarding explanation that would supposedly work by exhibiting naturalized rationality—one that mirrors the problems just discussed.

VI. Variations in competence, a second misgiving

Human rational competence is a select set of processes—among those processes that humans could undergo, some are jointly epistemically desirable, and only some are fitting or appropriate in some parallel sense that would make for practical rationality. Some of these select processes are presumably of a rudimentary sort—a natural cognitive endowment on which fuller rational competence can be developed. Some of these rudimentary components of competence may be operative pretty much from the beginning.[14] Others may take shape with normal development and little specific training.[15] Other components of rational competence would need to be acquired by way of systematic training. Certainly this is true of some of what makes for competence with probabilities. Humans can become more accurate and reliable in reasoning from information provided in several ways. When judging the probability of the joint occurrence of two independent features, A and B, they can employ processes that multiply, rather than crudely average, the estimated probabilities of A and of B. Further, if Gigerenzer and Hoffrage's (1995) recommendations are apt, humans might be taught to insist on formulating the certain classes of problems using an information format (the sampling-frequency format) that makes it relatively easy (for humans) to avoid the neglect of base rates that has been found to plague much reasoning in terms of probabilities. Arguably, such processes are then a part of full epistemic competence.

We can think of the set of tractable processes that might be humanly acquired with suitable training as a humanly fitting ideal. This ideal—this *ideal* human rational competence—is not matter of just some rudimentary subset of desirable human cognitive processes. Nor is it a matter of some subset that a particular pretty good human cognitive agent might come to have acquired. Ideal human rational competence is a matter of the full set of complementary tractable processes that humans could acquire (with training) and employ (with adequate motivation). This is a descriptively constrained ideal—constrained by what are the cognitive endowments and thus limits of humans as cognitive systems. For the naturalist, this is the full epistemic and rational competence that is necessarily connected with objective normative rationality—*human rationality is a matter of humanly ideal rational competence*.

Unfortunately, but predictably, all of us humans have failed to acquire full ideal human epistemic competence. What each actual human has attained is some proper subset of this select subset of tractable processes—that agent's *actual rational competence*.[16] For the naturalized rationalist, there may be no daylight between *ideal* rational competence and objective rationality (no possible divergence), but *there is clearly plenty of room for divergence between actual rational competence (on the one hand) and ideal rational competence or rationality (on the other)*. Once this is recognized, we see immediately that one can recast the argument advanced in the last section.

First, distinguish two central types of cases (parallel to those discussed earlier):

	The result is *weakly naturalistically rational*—it conforms to what could be gotten by *ideal* human competence
The result is gotten by the agent's *actual* rationally competent processes	1. The agent generalizes from a sample in a way that employs the agent's actual limited competence. This competence may be supposed to involve some trained sensitivity to possible biases. But, it may (typically will not) involve all the forms of general sensitivity to biases that humans might acquire.

The result is gotten by other processes	2. By wishful thinking the agent arrives at the belief that a certain feared combination of conditions is unlikely—as it is, the agent had on hand sample data that would have allowed that conclusion to be drawn using full ideal human competence. But, really, arriving at this conclusion was not sensitive to that data. Crudely similar data that would not have made for an ideally competent inference to that conclusion was used by less sensitive processes.

Table 2

One can then paraphrase the points made above in connection with aprioristic rationalizing explanation. Correct answers to what-if-thing-had-been-different questions, the truth about patterns of counterfactual dependencies, causal relevancies, are a function of the actual processes in play. If the processes in play are components of an *actual* rational competence that diverge significantly from full *ideal* human rational competence, *and if, as the naturalist must hold, such an actual competence falls short of objective rationality,* then the counterfactual dependencies between states will reflect the limitedly competent transitions rather than naturalized rational transitions associated with ideal competence. The answers to what-if-things-had-been-different-questions will reflect the fact that the agent's (or agents') conclusions or decisions were (and would generally be) competent *in a limited fashion*, the outcomes of a limited rational competence, but not generally objectively rational. In effect, in cases of type 1, the agent's or agents' actual limited rational competence screens-off naturalized rationality (associated with ideal rational competence) from explanatory relevance. Even in these cases, where the agent's actual rational competence is sufficient to mimic the results that would be gotten by ideal epistemic competence in a limited range of cases), naturalized rationality is not explanatorily relevant. (Compare: even in cases in which the red liquid in hand is blood, pouring blood into the water would explain the congregation of sharks, while pouring red liquid would not.)

Of course, explanation in terms of naturalized rationality in cases of type 2 has no plausibility. When processes in play are neither a component of actual or ideal epistemic

competence, the counterfactual dependencies between states will generally reflect neither competent transitions nor aprioristically rational transitions. Holding constant the processes here in play, the answers to what-if-things-had-been-different-questions will reflect the fact that the agent's (or agents') conclusions or decisions were (and would be) both rationally incompetent and generally irrational (naturalistically objectively irrational).

Thus, when focusing on the distinction between actual epistemic competence (on the one hand) and ideal epistemic competence and naturalized rationality (on the other hand), one finds a fatal flaw in the idea that it is explanatory to find an agent (or agents) to be naturalistically rational. The flaw is parallel to that identified in connection with the aprioristic conception of rationalizing explanation. As long as the naturalist about rationality recognizes distinctions corresponding to actual and ideal competence on the part of limited creatures such as we humans, the problem plaguing the apriorist identified in the preceding section has *mutatis mutandis* application to the naturalist as well as to the apriorist. Thus, while the naturalist may avoid one fatal objection to the idea that findings of rationality are explanatory, the naturalist yet faces a parallel objection. Naturalist rationality is no more explanatory of what is thought and done than putting red liquid in the water is explanatory of sharks congregating.

VII. Some positive suggestions and some cold water.

The discussion to this point need not be taken as wholly negative in what it suggests. Consider this suggestion: what might be explanatory, and what would at least *look like* explaining by exhibiting the rationality of what is thought or done, is better understood as explaining by exhibiting the result as one that could be gotten by way of the agent's actual competence. But this needs refinement in several respects. First, and most importantly, it is not sufficient for explanation that one appreciate that the result in question *could be* gotten by way of the agent's actual competence (or the agents' actual competences). This is because even an agent whose

competence would generally be up to forming the belief, preference, or decision in question may yet have come to the belief by way of other processes. In such cases, either the processes constituting the agent's competence were not in play, or their operation was interfered with by various inappropriate processes (which yet could have conspired together to yield the result that in this case happened to conform to the results that the agent's competence would have yielded). The result that conformed to the agent's actual rational competence would then have resulted from a performance error of some sort. If the explanandum is a performance error, the processes actually in play are not (or not only) components of the agent's actual competence. (The present points are parallel to those developed regarding weak rationality in the previous sections.)

Perhaps we can say that when an agent arrives at a result by way of an actual competence, without interference from other cognitive processes—so that there is no rational performance error—then that result's being the outcome of such actual (in play) competence seems explanatorily relevant. Making evident that this obtained would be explanatory. One might announce that in place of rationalizing explanation, we had found a place for *competent-ifying explanation* (although that phrase is likely too ugly find many takers).

Unfortunately, I doubt one can be so conciliatory. The idea that normative notions have an explanatory role must suffer one last indignity. All that matters in explanation turns on the processes that were in play—whether these were objectively rational or irrational is not of any relevance. (Of course the agent probably conceived of these as good or appropriate. But that does not make them objectively rational, or objective appropriate and competence constituting processes.) In any given episode, the processes in play were what they were. These may be given a descriptive abstract characterization in terms of inputs and outputs. This characterizes the aspects of the agent's cognitive processes that determine the patterns of counterfactual dependencies. The processes might be normatively appropriate, or not. Either way, this normative appropriateness seems itself screened off from being explanatorily relevant.

VIII. A clarification

I have relied on a rather generic notion of explanation—I have not relied on any of the "fancy" machinery of specific to certain contested lines of thought regarding explanation (the subsumptive, ontic or erotetic lines of thought, for example), and my arguments largely turns on what is common coin to all such accounts. Still, one might yet suspect that I have yet insisted on too much in the way I think about what it would be to reveal patterns of counterfactual dependencies.[17] I should clarify.

To begin with, I have said that reasons-explanations must afford an appreciation of how, were antecedent thoughts different, the resulting thought or action would likely have been different in certain ways. I think that we commonly do manage to satisfy this demand: reflect on some of the action explanations to which one would confidently commit oneself. To the extent that one can be confident that the agent acted or thought as they did because of certain antecedent beliefs and desires, one can say that, were the agent's then to have believed (desired) otherwise (think of some variations), they would have thought or acted in various others ways. Again, I believe we commonly are epistemically entitled to our confidence regarding these matters.

In addition, I have insisted that such patterns of counterfactual dependency turn on the cognitive processes in play in the agent whose thought or action is being explained: "Only these processes—the transitions to which they do and would give rise, the input on which they work in an episode, the features of that input to which those processes are sensitive—can be explanatory of what transpired in a given episode." Thus, reasons explanations turn on some such understanding of these processes.

Sometimes the character of those processes is articulated in the explanation itself. One might, for example, explicitly note that the agent employed a certain heuristic, or fast and frugal process, or that the highly trained agent at work at the National Institute of Health conscientiously calculated probabilities using Bayes' Theorem and drawing on antecedently established base

rates. Sometimes the processes will be only crudely alluded to, or left for one to surmise, as when one notes that the agent, with the help of free booze, had the pretty girl kiss the dice and then bet the farm. In such cases, I think, upon hearing of the booze and the seemingly superstitious behavior, one is (in effect) invited to draw on background information about the kinds of cognitive processes to which people are subject. In such cases, it is what is inferred regarding the processes in play that carries the explanatory weight. Commonly, one is given enough information to implicitly invite a kind of simulation of the other. In such cases, the information provided makes it plausible that there is a similarity between certain cognitive processes one can employ oneself and the processes in play in the agent.[18] (It may be added that the processes one then uses in the simulation may be processes that one yet considers inappropriate. One may have come to restrain such tendencies in one's own reasoning, and yet use those processes in simulations.) But, in all such cases, if the explanation is successful, one receiving it must come to understand the features of the agent's antecedent beliefs and desires to which the agent's processes are sensitive, and the corresponding patterns of counterfactual dependencies.

The information on which one relies, information regarding the processes in play in the episode, need not be perfectly accurate or precise. In the explanation of thoughts and actions it simply never is. But this is not unique to reasons explanations. In many explanations in many domains, one's understanding or characterization of the operative processes may be rough or approximate. Perhaps the molecules to which shark sensory organs differentially respond are just certain invariant components of blood. So, had we dumped these blood extracts into the water, the sharks would have congregated as they did. Still, it seems that the background information that sharks are attracted to blood gives us enough of an appreciation or understanding of patterns of counterfactual dependencies (in our hypothetical case) to provide a significant explanation. When I leave my basketball out in the sun, temperature of the contained air increases. The pressure of the gas also increases, and the ball volume increases very marginally. I might explain the increase in pressure using the ideal gas law—admittedly a crude generalization, but one with a reasonable

degree of invariance (Woodward 2000). It is not true of all gases across all temperature or pressure ranges that PV=rT, and yet this familiar generalization holds across a significant range of values for the parameters, and comfortably holds of the gaseous processes in my basketball episode. Thus, the ideal gas law allows us to appreciate or understand the pattern of counterfactual dependencies obtaining in this episode.

In view of the preceding paragraph, I cannot maintain that the problem with normative principles in explanations of what is actually done or thought in some episode is that they only approximately characterize the processes there in play. That would be true of information to the effect that the agent employed certain heuristics, or that the agent used such processes in a way suggested by certain background processes, or that the agent conscientiously employed a learned algorithm of some familiar sort. On reflection, one must admit that any characterization of cognitive processes in play in an agent will be approximate. At best, we have reasonably invariant generalizations—generalizations with limited but significant invariance. This is no less true when one takes one's explanatory clues form apparent cognitive similarities and simulates the other. (After all, we rely then on the idea of "similarity.") The problem with normative principles in explanations of why an agent actually acted or thought as the agent did must be different.

The problem, in a nutshell is that they are normative principles—and their correctness as normative not turn on what processes are in play in an episode. As normative principles, or as characterizations of what is normatively correct, they do not provide, or purport to provide, an explicit or a background understanding of processes in play in the case. In getting to an explanation there is always another step, and this further step screens of the normative status of processes as objectively rational from explanatory relevance.

Focus again on the pivotal class of cases: those in which the agent is weakly rational (in accordance with the relevant standards (aprioristic, ideal naturalized, or even those associated with the agent's actual competence, such as it is) and in which the processes in play in the agent are just components of the agent's actual rational competence. Suppose additionally the processes

in play are also components of the ideal human competence. Bracketing certain very fine points, we may suppose that there is some characterization of these processes that might do double duty: it is a descriptive characterization of the agent's processes and it characterizes processes that are normatively appropriate. Let *P* stand for this characterization. In cases of the pivotal sort, there will then be a cognitive scientific characterization of the agent's processes (as *P*), and a normative principle (humans Ought$^{\text{rationality}}$*P*). Of course, normative rationality commonly diverges from the processes in play. So that, one cannot generally infer from humans Ought$^{\text{rationality}}$*P* that an agent of concern had processes of kind *P* in play. But, now we are focusing on a class of cases in which it is true both that humans Ought$^{\text{rationality}}$*P* and that the agent of concern had processes of kind *P* in play.

One who is holding out hope for explanation by exhibiting rationality might think as follows. Look! *P* characterizes the processes in play, and it characterizes processes that ought to be in play. To explain the agent's resultant thought or act, we need only note that the agent was rational, or perhaps rational in the specific respect of honoring Ought$^{\text{rationality}}$*P,* and that the agent had the relevant input to *P.* Why is not that good enough?

There is a sense in which that is good enough. There is also a sense in which it may at least suggest too much. It certainly contains or suggests some explanatory information, but it also contains some information that is irrelevant to explanation. By stipulation, *P* characterizes both the processes in play and what processes ought to be in play. Thus, from the normative characterization of the case, we may suppose that one can infer that the process of type *P* was in play. This provides at least a rough characterization of the processes in play, and provides an (at least approximate) understanding of the pattern of counterfactual dependencies obtaining there. So, in the special class of cases on which we are focusing, *P* does double duty—featuring both in a descriptive characterization of what processes are in play there and in a normative characterization of what ought to be in play there. Given the normative characterization, and the information that the agent had the normatively approved processes in play, one obviously can

recover the descriptive characterization. There is no question that this affords an understanding of the patterns of counterfactual dependency then obtaining.

Compare: I may drop my compass at the top of K2, and it will accelerate at a certain rate. That rate will be different from what would have happened were I instead in Death Valley. From the point of view of what does the heavy lifting in the explanation of the precise acceleration of my compass, K2 is not special—its just tall and that is about all. It is something about the distance between the compass and the center of mass of the earth that is relevant. Of course, one might mention K2 in the explanation, but no one would think that being on K2 (however cool that might be) was explanatorily relevant. They would quickly note the indicated altitude and that would be taken to be what was explanatorily relevant—it would be what was important with respect to the processes in play. Being at 28,238 feet above sea level (or whatever that yields as distance from the center of the earth) is explanatorily relevant and screen off one's location on the peak of K2 from being relevant. (Being 840 feet below the peak of Mount Everest would not make a difference.)

To appreciate the pattern of counterfactual dependencies involving compasses, other items, and accelerations, one has to use information about one's geographical location to get to information about altitude, or distance from the center of the earth, and to use at least approximate information about how these parameters matter in the gravitational process in play. To appreciate the pattern of counterfactual dependencies involving beliefs, desires, and actions—even in those cases in which the agent is strongly competently rational—one has to use information about the normative propriety of the agent's process to get to at least a approximate descriptive characterization of the processes in play. One has to move from $\text{Ought}^{\text{rationality}}P$ and agent A thought as A ought, to A had processes P in play. Then the descriptive characterization of the processes in play does the explanatory work—enabling one to appreciate the patterns of counterfactual dependency obtaining in the episode.

The best way to understand much of the earlier argument is to see it as indicating ways in

which models of rationality are generally problematic vehicles for information about what processes are actually in play in an episode. There is much daylight between apriorist rationality and the processes that could be in play in humans. There is much daylight between ideal human competence and the processes in play in any actual human. This should not be surprising. For such normative models are not suited or intended to characterizing what processes are in play, but what processes ought to be in play.

---

[1] This paper benefited from the comments I received from audiances at the University of Alabama, Birmingham and Wichita State University, as well as at the 2009 Philosophy of Social Science Roundtable. Concerns pressed by Brian Epstein and Jeffery Hershfield (among others) prompted the final section of the paper. After presenting an earlier version of this paper at the roundtable, Stephen Turner shared with me his recent book manuscript, *Explaining Normativity*. We are, in his words, "locked on the same target." In his final chapter, Turner advances what may be the most helpful reading of Davidson I have encountered. If his reading is correct, then my objections to a Davidsonian position developed here do not apply to Davidson, although they apply to most "Davidsonians." He also makes particularly fecund use of material on simulation.

[2] Rosenberg's arguments depend on a (Davidsonian) view about rationality and interpretation. For critical discussion of Rosenberg's arguments, see Henderson (1993, 207) and Kincaid (1996).

[3] To write of certain kinds of cognitive processes being "expected" might be taken to suggest that explanation is here thought to depend on nomic or invariant generalizations characterizing cognitive processes in a class of systems, it might be taken as suggesting some subsumptive model of explanation. I do not favor such a limiting model. I think that we might also expect from other human agents kind of cognitive processes and transitions that we can readily simulate (Henderson 1996, Goldman 2006, Stueber 2006).

[4] This list of alternatives reflects Salmon's (1989) historical overview.

[5] While the points developed in this paper apply most directly to reasons-explanations of the thoughts and actions of individuals, they have wider application across the social sciences, insofar as those sciences treat of actions (or of the institutional effects of patterns of actions) and of beliefs (or information, or opinion), and of desires, (or preference, costs or benefits). This is to say that they apply to most all of the human sciences. If findings of rationality (or of findings that

the phenomena accords with what would be rational) are not explanatory in connection with individual agents and actions, they are not explanatory in connection with social or collective phenomena. This is not because all explanation must ultimately be given at the level of individuals and their actions. Nor is it because those social phenomena reduce to phenomena at the level of individuals. Such individualist doctrines seem far-fetched. Rather, it is because all explanations at the level of holistic entities—groups, institutions, and the like—presuppose that there are underlying mechanisms. It is because holistic phenomena supervene on lower level phenomena. This is not some a priori truth about explanation, but is rather a substantive point about how the world works. Talk of genes presupposes that there are biochemical entities whose causal interactions constitute the genes and their assortment, recombination, transcription and expression. One can explain the distribution of some trait in successive generations by filling in the story at the level of genes, without giving or even knowing the story at the level of biochemical realizations. But, were there no such phenomena at that lower level (perhaps one allowing for moderate correction or refinement), then any story at the level of genes would be an unexplanatory myth.

[6] The naturalized epistemological approach, or the naturalized approach to rationality more generally, would have a place for a parallel set of distinctions, but would need to conceive of the contrast somewhat differently.

[7] Arguably, just as talk of *agents* being rational is subject to a weak and a strong reading, so is common talk of *beliefs* being rational. I mentioned only the weak reading above. Let that serve as a stipulation regarding my use of such talk in this paper.

[8] Some qualification is necessary—because notions of grammatical correctness is itself somewhat varied. To start with, one should draw a distinction between two senses of grammatically correct performance—for one finds here an analog of the distinction between weak and strong rationality. There is a sense in which one's grammatical performance is correct—in a given language or dialect—if it is formally grammatical. This might be termed weak grammatical correctness. This turns on a formal notion of grammatical correctness that applies to strings of phonemes, however long. It characterizes grammatical sentences, not primarily linguistic performances. Alternatively there is a sense in which one's linguistic performance is correct only if it is generated through one's grammatical competence. This is the analog of strong rationality. I might recognize as grammatical a sentence in Hindi by laboriously checking it against an inscription on a list of Hindi sentences. My verdict that the Hindi sentence is grammatical does not count as correct

performance in Hindi, even though my verdict regarding it is correct (and perhaps a correct performance in English). My reading the sentence off of the list would not count as a grammatically correct performance in the strong sense—although it would count as grammatically correct in the weak sense. Compare: my verdict that some complex formula is a logical truth by virtue of laboriously checking it against an inscription on some unlabeled list some stranger just handed to me need not count as a case of rationality—strong rationality—even though that formula is a logical truth and my thinking it true is thus weakly rational. (Compare these thoughts to those regarding the correctness of the child's fielding competence to come in the paper.)

[9] The divergence envisioned here between apriorist rationality and competence is not merely a matter of the limits of human memory. That is, it is not that the transitions to which competence would give rise would be exactly those that the apriorist would demand were it not for the bottleneck resulting from limited working memory. Some processes that constitute parts of human epistemic competence are likely to be heuristics whose use is marginally conditioned by some cues to when these are not likely unreliable. Some kinds of aprioristically called for transitions may be difficult for humans, so that they can only be managed, if at all, in a very limited way. One possibility along these lines is suggested by Gigerenzer and Hoffrage (1995), who argue that information provided in terms of probabilities, rather than in terms of parallel sampling frequencies, may be difficult for human agents to process in ways that conform to aprioristic demands. Cherniak (1986) suggested that there might be robust regularities in the differential feasibility for humans of various forms of inferences—important here are differences across those forms of inference which might be aprioristically recommended. Human epistemic competence might involve not engaging in courses of belief-fixation that rely heavily on inferential transformations that themselves are low on the human feasibility ordering. This restriction of competence constituting processes could easily result in humans foregoing certain inferential processes that would yet be called for aprioristically. It might also call for "substitute processes" that only approximate in a limited way what would be aprioristically demanded—it might call for heuristics. This would obtain even on those the operation of the heuristics best disciplined or conditioned by wider cognitive processes as in many "two systems" approaches (Sloman 2006, Kahneman and Frederick 2005).

[10] There is a sense in which there is wide agreement on this point. At the same time, one must acknowledge that different writers would develop the suggestion in very different ways. Those

who have worked in the heuristics and biases tradition—growing out of work such as Kahneman D. and Tversky A. (1972) and Kahneman D., Slovic, P., and Tversky, A. (1982)—would think of the heuristics needing significant discipline from limited ranges of more traditional processes. (Kahneman and Frederick, 2005, provides a recent formulation.). Gigerenzer and his collaborators (Gigerenzer and Hoffrage, 1995; Gigerenzer and Selton, 2002) champion fast and frugal processes said to be part of an adaptive toolbox with significant "ecological validity." They would see comparatively little need for these processes being controlled by more traditionally demanded processes. Johnson-Laird and his collaborators (Johnson-Laird, 2006, Johnson-Laird and Byrne, 2002; Johnson-Laird, Legrenzi, Girotto, Legrenzi, and Caverni, 1999) provide a somewhat more equivocal picture in which one uses mental models. While use of these could stand for some refinement, they would remain an ineliminable component of competence, and would not fully conform to apriorist demands. See also Bishop and Trout (2006). Some of the relevant literature is surveyed in recent review articles focused on whether one can best understand human cognitive processes as comprised of two systems (a system of formal algorithmic processes that polices system of fast heuristic processes), or rather is best understood as a less hierarchical assemblage of processes of various kinds (Sloman 1996, Osman 2004). See also Samuels, Stich, and Bishop (2002).

[11] It should be clear that both naturalized epistemology and rationality naturalized, as understood here, are methodological positions rather than metaphysical positions. They do not seek to make epistemological value, or any value, naturalistically or physicalistically respectable (by some reduction or by some show of supervenience, for example). Rather, they seek scientifically informed, realistic, epistemic standards.

[12] If one thinks that apriorist rationality makes different demands, the examples would need modification, but as long as there is divergence between what strong rationality requires and competent performance, there will be examples of each sort.

[13] Suppose the agent is provided information (from large studies) about would allow complex Baysian calculations of the exact probability of having some medical condition given the positive result on the test one just received. Then suppose that one also is given actual sampling frequency data regarding a folk in one's local health provider's office (somewhat smaller samples) of folk with positive results, negative results, and who turned out to have or not have the condition in question. Gigerenzer and Hoffrage (1995) argue that information in a sampling frequency format makes it far more tractable for people to, in effect, accommodate the base rates. So, the agent

might be faced with an intractable or nearly intractable problem dealing with a wonderful data set (provided), or a much more tractable problem, but one that would have the agent reason from what is a much smaller sample. Apriorist principles would seem to call for the fancy reasoning drawing on the full information one is provided. A reasonable understanding of human competence might call for the much more tractable reasoning drawing on the more limited data set. So, suppose that the agent competently comes to a value that is different from the fancy value. (The method Gigerenzer and Hoffrage discuss is not technically a heuristic, and is rather a simple special purpose process, something on the order of an limited algorithm that is yet a part of Gigerenzer (2002) would term our adaptive toolbox. If one requires the use of the best information possessed, then the case envisioned is falling into cell 3. If one does not require the use of the full information possessed, it might be treated as falling into

A different and interesting borderline case, one that might be categorized in either cell 1 or 2 is this: Using processes that satisfice, where these cannot be known by the agent to maximize, the agent happens to hit upon a choice that is a maxima. Presumably because the result is a maxima, it is at least weakly aprioristically rational. But, aprioristic principles are commonly thought to require that one satisfice only where that method is itself a maxima, where it is likely that the results of further search and inquiry would not bring marginal improvements in the value of verdicts that justify the information costs. Supposing that the agent is not in a position to know this, the choice of a stopping place would commonly be thought to be an instance of type 2.

[14] For example, there are apparently perceptual mechanism that facilitate the integration of partially occluded objects, and that dispose the enfant to find object continuity across presentations. (Wolf, et. al., 2006, chapter 4 summarizes some of the relevant results.

[15] For example, work in developmental psychology having to do with when children typically acquire a capacity for imaginative simulation adequate to false belief tasks suggests that these refinements in human'

[16] Actually, this may be an overly simplistic characterization of one's actual cognitive competence. It may include processes that would not have a place in the more ideal human competence, which is, in effect a kind of human optima. It is plausible that an agent's actual competence may be constituted by a set of processes that jointly satisfice, but are not jointly optimal. Such a set may include processes that are not a member of the optimal human set of processes.

[17] Certainly talk of explanation can be pretty heterogeneous. This is because the generic concept is best understood erotetically—as a matter of answers to questions. One can explain why one

ought to do something—giving an answer to the normative question. Here rationality, or other normative matters, might be "explanatorily relevant." The explanations of concern here are those that explain why an agent (or agents) thought or acted as they actually did. Whatever one's detailed story about what answers to this question require, the idea of revealing or exhibiting patterns of counterfactual dependencies will play a central role.

[18] Again, nothing I have said here rules out explanations in which, rather than drawing on descriptions of the agent's processes, one simulates those processes (Steuber 2006; Goldman, 2006, Turner MS).

References

Cummins, R. 1983. *The Nature of Psychological Explanation.* Cambridge, MA: M.I.T. Press.

Davidson, D. 1980a. Mental Events. In his *Essays on Actions and Events.* Oxford: Clarendon Press: 207–225.

Davidson, D. 1980b. Towards a unified theory of meaning and action. *Grazer Philosophical Studien* 2: 1–12.

Bishop, M. and Trout, J. 2005. *Epistemology and the Psychology of Human Judgment.* Oxford University Press.

Davidson, D. 1982. Paradoxes of Irrationality. In *Philosophical Essays on Freud,* edited by R. Wollheim and J. Hopkins*.* Cambridge: Cambridge University Press: 289-305.

Chernaik, C. 1986. *Minimal Rationality*. M.I.T. Press.

Gigerenzer, G and Hoffrage, U. 1995. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review* 102: 684-704.

Gigerenzer, G. and Selten, R. (eds.) 2002. *Bounded Rationality: The Adaptive Toolbox.* Cambridge, MA: M.I.T Press.

Goldman, A. 1976. What Is Justified Belief. In *Justification and Knowledge*, edited by G. S. Pappas. Dordrecht: D. Reidel: ??

Goldman, A. 1986. *Epistemology and Cognition*. Cambridge: Harvard University Press.

Goldman, A. 1992a. Strong and Weak Justification. In his *Liaisons*. Cambridge, MA: M.I.T. Press: 127-141.

Goldman, A. 1992b. Epistemic Folkways and Scientific Epistemology. In his *Liaisons.* Cambridge, MA: M.I.T. Press: 143-175.

Goldman, A. 1999. A Priori Warrant and Naturalistic Epistemology. *Philosophical Perspectives* 13: 1-28.

Goldman, A. 2006. *Simulating Minds.* New York: Oxford University Press.

Henderson, D. 1993. *Interpretation and Explanation in the Human Sciences*. Albany: SUNY Press.

Henderson, D. 1996. Simulation Theory vs. Simulation Theory: A Difference Without a Difference in Explanation. *Southern Journal of Philosophy, Spindel Conference Supplement* 34: 65-94.

Henderson, D. 2002. Norms, Normative Principles, and Explanation. *Philosophy of Social Science* 32: 329-364.

Henderson, D. 2007. Rationality and Rationalist Approaches in the Social Sciences. In *Handbook of Social Science Methodology*, edited by S. Turner and W. Outwaite. Sage Publishing: 282-301.

Henderson, D. 2008. Testimonial Belief and Epistemic Competence. *Nous* 42: 190-221

Henderson, D. and Horgan, T. 2008a. Would You Really Rather Be Lucky Than Good? On the Normative Status of Naturalizing Epistemology. In *Naturalism, Reference and*

Ontology: Essays in Honor of Roger F. Gibson, edited by Chase Wrenn. New York: Lang Publishing:47-75.

Johnson-Laird, P. 2006. *How We Reason.* Oxford: Oxford University Press.

Johnson-Laird, P. and Byrne R. 2002. Conditionals: A Theory of Meaning, Pragmatics, and Inference. *Psychological Review* 109: 648-678.

Johnson-Laird, P., Legrenzi, P., Girotto, V., Legrenzi, M., and Caverni, J. 1999. Naïve probability: A mental model theory of extensional reasoning. *Acta Psychologicaq* 93: 62-88.

Kahneman, D. and Frederick, S. 2005. A Model of Heuristic Judgment. In *The Cambridge Handbook of Thinking and Reasoning,* edited by K. Holyoak and R. Morrison. Cambridge: Cambridge University Press, pp. 267-293.

Kahneman D. and Tversky A. 1972, Subjective Probability: A Judgment of Representativeness.*Cognitive Psychology* 3: 430-454.

Kahneman D., Slovic, P., and Tversky, A. 1982. *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press.

Kincaid H. 1996. *Philosophical Foundations of the Social Sciences.* Cambridge: Cambridge University Press.

Osman, M. 2004. An Evaluation of Dual-Process Theories of Reasoning. *Psychonomic Bulletin and Review* 11: 988-1010.

Quine, W. 1969. Epistemology Naturalized. In his *Ontological Relativity and Other Essays.* New York: Colombia Univ. Press: 69-90.

Quine, W. 1986. Reply to Morton White. In *The Philosophy of W. V. Quine*, edited by L. Hahn and P. Schilpp. La Salle, IL: Open Court: 663-665.

Salmon, W. 1989. Four Decades of Scientific Explanation. In *Scientific Explanation, Minnesota Studies in the Philosophy of Science, vol. 13*, edited by Kitcher, P. and Salmon, W. Minneapolis: University of Minnesota Press: 2-219.

Samuels, R., Stich, S., and Bishop, M. 2002. Ending the Rationality Wars: How to Make Disputes about Human Rationality Disappear. In *Common Sense, Reasoning, and Rationality*, edited by R. Elio. New York: Oxford University Press: 236-268.

Sloman, S. 1996. The Empirical Case for Two Systems of Reasoning. *Psychological Bulletin* 119: 3-22.

Steuber, K. 2006. *Rediscovering Empathy*. Cambridge, MA: M.I.T. Press.

Turner, S. Manuscript. *Explaining the Normative*.

Wolf, J., Klunder, K, and Levi, D. 2006. *Sensation and Perception*. Sunderland, MA: Sinauer Associates.

Woodward, J. 2000. Explanation and Invariance in the Special Sciences. *British Journal for Philosophy of Science* 51: 197-254.