Philosophy and
Phenomenological Research

Philosophy and Phenomenological Research
Vol. LXXXI No. 1, July 2010

Higher-Order Evidence[1]

DAVID CHRISTENSEN
Brown University

Any new evidence that's worth its salt—that is, any evidence that rationalizes a change of belief state—is, in a trivial way, evidence against one's previous belief state. If I get an updated weather forecast, I may be rationally required to decrease my credence in sun tomorrow because my old credence in sun is not the appropriate one, given the meteorological evidence I now have. But while this sort of evidence does indicate that my previous beliefs are, in a certain sense, suboptimal, it does not indicate that I've been anything less than a perfectly rational believer. The evidence that there's something suboptimal about my pre-change beliefs is merely a byproduct of the evidence bearing directly on the subject matter of the beliefs.

Sometimes, however, evidence rationalizes a change of belief precisely because it indicates that my former beliefs were rationally sub-par. This is evidence of my own rational failure. If I learn that I've been systematically too optimistic in my weather predictions, I may also be rationally required to decrease my credence in fair weather tomorrow. But in this case, the indication that my former beliefs are suboptimal is no mere byproduct of my reasoning about the weather. What I learn bears on meteorological matters only via indicating my rational failings; intuitively, one might even balk at thinking of information about my optimistic tendencies as "evidence about the weather".

These two ways that evidence rationalizes change of belief correspond to two ways in which I'm a fallible thinker. One dimension of my fallibility is that my beliefs are based on limited evidence. So the conclusions I draw—no matter how competently I react to my

[1] I'd like to thank Nathan Ballantyne, Adam Elga, Ram Neta, Josh Schechter, Jonathan Vogel, Jonathan Weisberg, and the students in my seminar at Brown for valuable written comments or conversations about the topics discussed here. This paper was discussed at Epistemology Reading Groups at Brown and MIT, at Tom Kelly's seminar at Princeton, and at the Rutgers Epistemology Conference; many thanks to the discussants at all four occasions.

evidence—may turn out to be inaccurate. Recognition of this sort of fallibility occasions doubts of the underdetermination variety. The second dimension of my fallibility is that I may make mistakes in thinking. I sometimes fail to draw the conclusions that are best supported by my evidence, often because I make mistakes in judging what my evidence supports. The doubts occasioned by recognition of this second sort of fallibility seem, at least at first, to be different from doubts based on underdetermination. The two bits of evidence described above rationalize belief-revision by addressing, respectively, these two sorts of fallibility.

The apparent contrast between these two roles that evidence can play prompts the question of how deep the difference between them really lies. Most discussion of evidence has clearly focused on the first sort of paradigm. But if the second role is really different, that would leave open the possibility that thinking about the second sort of paradigm may reveal distinctive features of rationality. That is the possibility I want to examine below. Let us begin by looking at some examples of what, following Thomas Kelly, I'll call "higher-order-evidence" (HOE). (Richard Feldman (2005, 100) has called it "second-order evidence," though some of his examples differ from the paradigms I'll be focusing on.) Kelly and Feldman characterize their notion as evidence about evidential relations; the idea is that when I get evidence of my own epistemic malfunction, this serves as evidence that the evidential relations may not be as I've taken them to be. I won't try to give a precise characterization of HOE, but will instead work with some paradigmatic cases of evidence suggesting my own epistemic failure, in order to consider some of the ways that this sort of evidence may be distinctive.

1. Some Examples of Higher-Order Evidence

a. Reasonable Prudence

I'm a medical resident who diagnoses patients and prescribes appropriate treatment. After diagnosing a particular patient's condition and prescribing certain medications, I'm informed by a nurse that I've been awake for 36 hours. Knowing what I do about people's propensities to make cognitive errors when sleep-deprived (or perhaps even knowing my own poor diagnostic track-record under such circumstances), I reduce my confidence in my diagnosis and prescription, pending a careful recheck of my thinking.

b. Peer Disagreement

My friend and I have been going out to dinner for many years. We always tip 20% and divide the bill equally, and we always do the math

in our heads. We're quite accurate, but on those occasions where we've disagreed in the past, we've been right equally often. This evening seems typical, in that I don't feel unusually tired or alert, and neither my friend nor I have had more wine or coffee than usual. I get $43 in my mental calculation, and become quite confident of this answer. But then my friend says she got $45. I dramatically reduce my confidence that $43 is the right answer, and dramatically increase my confidence that $45 is correct, to the point that I have roughly equal confidence in each of the two answers.

c. Drugs

I'm asked to be a subject in an experiment. Subjects are given a drug, and then asked to draw conclusions about simple logical puzzles. The drug has been shown to degrade people's performance in just this type of task quite sharply. In fact, the 80% of people who are susceptible to the drug can understand the parameters of the puzzles clearly, but their logic-puzzle reasoning is so impaired that they almost invariably come up with the wrong answers. Interestingly, the drug leaves people feeling quite normal, and they don't notice any impairment. In fact, I'm shown videos of subjects expressing extreme confidence in the patently absurd claims they're making about puzzle questions. This sounds like fun, so I accept the offer, and, after sipping a coffee while reading the consent form, I tell them I'm ready to begin. Before giving me any pills, they give me a practice question:

> Suppose that all bulls are fierce and Ferdinand is not a fierce bull. Which of the following must be true? (a) Ferdinand is fierce; (b) Ferdinand is not fierce; (c) Ferdinand is a bull; (d) Ferdinand is not a bull.

I become extremely confident that the answer is that only (d) must be true. But then I'm told that the coffee they gave me actually was laced with the drug. My confidence that the answer is "only (d)" drops dramatically.

d. Anti-expertise Paradox

I'm a neurologist, and know there's a device that has been shown to induce the following state in people: they believe that their brains are in state S iff their brains are not in state S. I watch many trials with the device, and become extremely confident that it's extremely reliable. I'm also confident that my brain is not in state S. Then the device is placed on my head and switched on. My confidence that

my brain is not in state S ..... well, it's not clear here what should happen here.

All of these examples involve my being confronted with evidence that suggests my epistemic failure. In Reasonable Prudence and Drugs, the evidence is direct: the sleep-deprivation and drugs are taken as likely causes of epistemic malfunction. In Peer Disagreement, the evidence is both less pure and less direct. After all, my peer's belief is, in part, just evidence that my share is $45. But my friend's disagreeing with me, when we have exactly the same evidence, also strongly supports the hypothesis that one of us has made a epistemic mistake. And our extensive and equally good track records support giving significant credence to the mistake being mine. Finally, the Anti-expertise Paradox case involves evidence that I will not form beliefs about my brain-state in a reliable way.

2. Is HOE Really "Evidence"?

In thinking about examples of HOE, it's tempting to wonder whether what we're thinking about is really evidence at all—or, more precisely, whether it's evidence relevant to the propositions that are the subject of the affected belief. Why should we think, for example, that the amount of time since I last slept is relevant to whether some particular treatment is appropriate for a certain patient?

Of course, it's obvious that much evidence bears only indirectly on the propositions to which it's relevant. I may find out that a yellow Hummer was seen driving down a murder-victim's street shortly after the murder. This may give me evidence that Jocko committed the murder—but only because of my background belief that Jocko drives a yellow Hummer. This background belief is needed to make the connection between the Hummer-sighting and the murderer's identity. But HOE can seem like non-evidence in a sense that isn't just due to indirectness. While the information about what Jocko drives is essential to the bearing of the Hummer-sighting on the question of who committed the murder, no fact about my being well-rested seems to be needed in order for the basic symptoms my patient exhibits to bear on the question of what the best treatment is for her.

Kelly (2005), writing about the opinions of others as HOE, makes the following observation, which highlights an oddity of taking HOE to be just more evidence. Suppose that two people have shared first-order evidence E which bears on proposition P. Alice forms the belief that P on the basis of E before Ben comes to any judgment on the matter, and she tells Ben her opinion. Should Ben now take this information as additional evidence supporting P, over and above that provided

by E? Kelly points out that we don't think that *Alice* should consider P supported by both E and "I believe P on the basis of E". And if Alice doesn't take her own belief as additional evidence supporting P, it seems odd for Ben to take Alice's belief as additional evidence supporting P.

Hartry Field (2000), in defending strong apriorism—the thesis that a priori justification is empirically indefeasible—offers an account of evidence that would seem to exclude HOE from being evidence in a principled way. Field defines evidence in terms of "ideal credibility," which ignores computational limitations. Field takes logical truths, for example, to be ideally fully credible. So something that might seem to be empirical evidence defeating one's justification for belief in a logical truth (for example, the fact that some famous logicians disbelieve it, which suggests that one has made a mistake in one's own thinking) is not, strictly speaking, evidence bearing on the relevant proposition at all.

I think that neither of these points should be taken as showing that HOE isn't really evidence. Indeed, neither author takes his point to constitute a strong general argument against taking HOE as evidence.[2] But each author highlights an interesting feature of HOE that's worth making explicit.

Let's look first at Kelly's observation. Alice and Ben in a clear sense share the evidence (E, and the fact that Alice believes P on the basis of E). But it seems that, insofar as we countenance taking the higher-order part of the evidence as relevant to P, this would only be true for Ben! What should we make of this?

I think that what Kelly is drawing our attention to is a somewhat unusual feature of HOE in general: its evidential bearing is often relative to the thinker. A more stark example is provided by the Drugs case described above. My friend and I may share all the same information about the Ferdinand puzzle. Yet the information that I've been slipped the reason-distorting drug, which is information that she and I share, seems to have a dramatic bearing on what I should believe about the puzzle, but no bearing at all on what my friend should believe.

In each of these cases, the reason for agent-relativity is not mysterious. In the Drugs case, the reason is transparent. But even in the Alice and Ben case, once one focuses on why Alice's belief might bear on how confident Ben should be about P, the relativity makes sense. Alice's having formed the belief that P on the basis of E is evidence for Ben that E supports P, in the following way: If Ben is otherwise inclined to believe not-P on the basis of E, the HOE about Alice's

[2] In fact, Kelly (forthcoming) explicitly takes HOE to be evidence.

belief should raise worries that he's mistaken the bearing of E, and he should be less confident in not-P (and more confident in P). Alice's thinking thus serves for Ben as a check on his own. But it obviously cannot do that for Alice. So for Alice, this item of HOE should not affect her confidence in P.

Still, I think that Kelly was entirely correct in thinking there is something odd about this sort of agent-relativity. It seems to go against a very natural picture of evidence. On the natural picture, the import of a particular bit of evidence may depend on the thinker's background beliefs, but it does not depend on who the thinker is. But while this picture may typically be correct for the sort of evidence philosophers usually think about, HOE seems often to produce cases where the import of a certain bit of evidence varies considerably for agents who would be naturally described as sharing the same background beliefs. One might, of course, preserve a version of the natural picture. One could do this by simply admitting indexical propositions such as "I am Alice" as part of the total evidence. I have no reason to think this would be a bad thing to do; in fact, it would allow for the influence of an agent's intermediate degrees of confidence that, e.g., she was Alice. But preserving the natural picture in this way would simply recognize, not erase, the aspect of HOE that we've been examining. For admitting these sorts of indexical propositions as evidence would be motivated by the realization that HOE often makes information about the identity of the thinker relevant to the import of parts of her evidence.[3]

Let us turn now to Field's (2000) proposal: that evidence, strictly construed, should be understood as that which bears on a proposition's credibility once one ignores the agent's computational limitations. It was advanced to prevent empirical propositions from counting as defeating evidence bearing on a priori propositions such as logical truths, the assumption being that, ideally, logical truths would be maximally credible. Since empirical propositions (e.g., ones involving disbelief by famous logicians) would not lower the *ideal* credibility of a logical truth, they would not count as evidence bearing on that logical truth.[4]

---

[3] I do not wish to argue that only HOE ever has this agent-relative import. My point is that the agent-relativity of HOE can reinforce the impression that it's in some sense not proper evidence relevant to the propositions on which it bears in an agent-relative way. Thanks to Josh Schechter for pressing this point.

[4] I should emphasize that Field does not want to deny that agents such as ourselves should lose confidence in complex logical truths when experts disagree with us. His point is to capture a sense of "evidence" on which empirical evidence cannot undermine a priori justification.

To the extent that this proposal would bar HOE from bearing on logical truths, it would seem also to exclude HOE from being evidence in other sorts of cases as well. Consider empirical propositions (such as the one describing my friend's opinion in the Peer Disagreement case) which cast doubt on my empirical beliefs by casting doubt on the mathematical reasoning on which they were based. If ignoring computational limitations sufficed to make the ideal credibility of a logical truth immune from higher-order undermining, the same would presumably apply to the reasoning in the restaurant. So the disagreement of my friend would presumably not be counted as evidence bearing on what my share of the bill is.

And the point might well extend even to inductively supported conclusions. There's not obviously a motivated difference between the sort of fallibility that prevents ordinary agents from faultlessly appreciating logical relations, and the sort of fallibility that prevents ordinary agents from faultlessly grasping inductive support relations. So on Field's proposal, we might exclude from the realm of evidence proper any information that seems to affect an agent's rational credence in a proposition by bearing on the possibility that the agent has failed to grasp evidential relations perfectly.[5]

Now I do have some reservations about this notion of ideal credibility. Let us consider for a minute an agent who is, in fact, computationally unlimited, and let us grant that this includes not only perfect grasp of deductive logical relations, but also perfect grasp of inductive support relations. (It seems to me that we're not only idealizing away from what one would ordinarily think of as computational limitations here, but also supposing something further: that our agent correctly takes the deductive and inductive support relations for what they are. But I take it that this is in the spirit of Field's proposal.) Let us call such an agent "cognitively perfect".

Could HOE affect what such a cognitively perfect agent may rationally believe? It seems to me that it could. After all, even a cognitively perfect agent may well get powerful evidence that she is less than perfect. She might get powerful evidence, for example, that she'd been dosed with a reason-distorting drug that gave its victims the strong illusion of seeing clearly and distinctly the truth of claims that were in fact false. Of course, this evidence would, in the present case, be misleading evidence. But misleading evidence is exactly evidence that leads *rational* belief away from the truth; cognitively perfect agents will in general respect misleading evidence scrupulously. And I don't see how the mere fact of our agent's cognitive perfection

[5] Field himself intends at least our basic inductive methods to be a priori.

would make it rational for her simply to disregard the misleading evidence in this case.

For this reason, it seems to me that ideal credibility, if it's defined to mean credibility independent of considerations of possible cognitive imperfection, is not the same as rational credibility for a cognitively perfect agent. So we cannot exclude HOE from the realm of evidence on the ground that it's not the kind of thing that would affect the beliefs of an agent who always responded maximally rationally to her evidence.

Of course, one could stipulatively define "ideal credibility" as credibility independent of worries about cognitive imperfection. But that would give up the appeal that the more intuitive notion of ideal credibility might lend to a strong account of apriority. Field's more recent (2005) account instead bypasses the whole issue of ideal credibility. In order to rule out evidence such as the logician's disagreement as possibly undermining the justification of logical truths, he opts for simply ruling such evidence out as not counting, and goes on to sketch a tentative explanation of why this sort of evidence shouldn't count:

> A rough stab at explaining why they shouldn't count–doubtless inadequate–is to put the empirical unrevisability requirement as follows: there is no possible empirical evidence against p which is "direct" as opposed to going via evidence of the reliability or unreliability of those who believe or disbelieve p. (2005, 71)

This sort of explanation, of course, essentially just rules out HOE *per se* as counting for the purpose of defining apriority.

I would make two observations prompted by thinking about Field's proposals. Most obviously, even if one rejects the view that ideal credibility is unaffected by HOE, it does seem that HOE is distinctive in a somewhat milder way. The idea that a priori justification is empirically indefeasible is of course contested, but it's certainly an idea with considerable history and attraction. Yet it seems that fleshing out the idea requires segregating HOE from ordinary empirical evidence. This suggests that HOE is works in a way that's interestingly different from the way most empirical evidence works.

The second observation is that there is something right about the suggestion that HOE is irrelevant to ideal credibility, even if we think that cognitively perfect thinkers cannot rationally ignore HOE. Consider the case of a cognitively perfect agent who initially comes to believe P (which is, of course, exactly the correct belief, given her evidence). P might be a logical truth, or it might be supported by Inference to the Best Explanation (or whatever the correct inductive

rule is[6]) applied to her ordinary evidence. She then encounters strong HOE indicating that, due to cognitive malfunction, she has become too confident in P; in response, she reduces her confidence in P. Even if the reduction in her confidence in P is rationally required, we can see that, from a certain point of view, or in a certain dimension, her epistemic position has worsened. Even though she's capable of perfect logical insight, and even if she flawlessly appreciates which hypotheses best explain the evidence she has, she cannot form the beliefs best supported by those logical or explanatory relations that she fully grasps. So it seems to me that resistance to seeing HOE as relevant to ideal credibility may flow from the fact that, in taking HOE into account in certain cases, an agent is forced to embody a kind of epistemic imperfection.[7]

So, to sum up: HOE really is best thought of as evidence. It is information that affects what beliefs an agent (even an ideal agent) is epistemically rational in forming. But it seems, at first blush, to be evidence of a peculiar sort. For one thing, its evidential import is often agent-relative. For another, respecting it can apparently force an agent to fall short in certain ways, by having beliefs that fail to respect logic or basic inductive support relations. In the next section, I'd like to look more carefully at the apparent peculiarity of HOE.

3. How Does HOE-based Undermining Compare to Ordinary Undermining?

The examples of HOE that we've been concentrating on, and that I want to continue concentrating on, might naturally be thought of as *undermining* evidence. So a natural question, in studying the apparent peculiarities of HOE, is how the undermining in our HOE examples might compare to more standard examples of undermining evidence.

John Pollock (1986) emphasizes the distinctive nature of what he calls "undercutting defeaters". He notes that while the simplest way for the justification for a belief to be defeated is by evidence for its negation, another important sort of defeater "attacks the connection between the evidence and the conclusion, rather than attacking the conclusion itself" (39). I should note that on Pollock's view, deductive reasons (such as the ones apparently undermined in the Drugs example) are never subject to undercutting defeat (45). But perhaps we may put

---

[6] I'll generally use IBE below to stand for whatever basic inductive rule is correct; I don't believe that my arguments will depend on this choice.

[7] The connection between self-doubt and epistemic imperfection, especially as it involves ideal agents, is explored in Christensen (2007b). Much of what follows here continues this line of thought, with particular emphasis on non-ideal thinkers. Also, here I'll concentrate on self-doubt that's occasioned by HOE, and ignore complexities brought up by self-doubts not occasioned by HOE.

this part of Pollock's view aside in seeing how HOE-based defeat compares to the sort of undercutting defeat described by Pollock.

Richard Feldman (2005, 111–113) has noted that when my belief is undermined by HOE, the undermining seems different from the standard cases of undercutting defeaters. Feldman first notes the similarity: in both cases, the new evidence in some sense attacks the connection between the evidence and the conclusion. But he then notes that there seems to be a real difference.

Feldman considers a typical example of undercutting defeat: the justification for my belief that an object is red, on the basis of its looking red, is undercut by the information that it's illuminated by red lights. Feldman points out that the undercutting leaves intact the general connection between appearing red and being red; in fact, Feldman holds that the object's appearance remains a reason for believing it to be red. The defeater just shows that this reason cannot be relied on in the present case. By contrast, HOE seems to attack the general connection between evidence and hypothesis: "it [defeats] not by claiming that a commonly present connection fails to hold in a particular case, but rather by denying that there is an evidential connection at all" (2005, 113).

While I think there is something right about this suggestion, which does distinguish the red light case from, e.g., the Drugs case, I don't think it fully captures what's different about HOE defeaters.[8] Consider a case where justification proceeds via an empirically supported background belief, as in the case where the sighting of a yellow Hummer on the murder-victim's street supports the hypothesis that Jocko is the murderer. My justification can be undercut by my finding out that Jocko's Hummer was repossessed a month before the crime, and he's now driving a used beige Hyundai. But this case of ordinary undercutting does not leave intact any general connection between yellow-Hummer-sightings at crime scenes and Jocko's guilt. So I agree with Feldman's observation that there is a difference, but I suspect that the difference must be explained in another way.

I think that one way of putting our finger on the difference comes out when we consider the Drugs example above. There, I become much less confident of my belief about the correct answer to the logic puzzle when I'm told that I've been drugged. And this seems like the rational response for me to make, even if, as it turns out, I happen to be one of

[8] I should note that Feldman is not primarily concerned to describe the difference between HOE-based defeat and ordinary undercutting, but rather to argue that one can't dismiss HOE defeat by distinguishing it from ordinary undercutting. The present paper is obviously fully in accord with Feldman's main point.

the lucky 20% who are immune to the drug, and my original reasoning was flawless.

If you doubt that my confidence should be reduced, ask yourself whether I'd be reasonable in betting heavily on the correctness of my answer. Or consider the variant where my conclusion concerns the drug dosage for a critical patient, and ask yourself if it would be morally acceptable for me to write the prescription without getting someone else to corroborate my judgment. Insofar as I'm morally obliged to corroborate, it's because the information about my being drugged should lower my confidence in my conclusion.

If this is right, it seems to me to point to a different way of characterizing the peculiarity of HOE-based undermining. For in the case where I'm immune, it is not obvious why my total evidence, after I learn about the drug, does not support my original conclusion just as strongly as it did beforehand. After all, the parameters of the puzzle are not rendered doubtful by my new information. The undermining is directed only at the simple deductive reasoning connecting these parameters to my answer. So there is a clear sense in which the facts which are not in doubt—the parameters of the puzzle—leave no room for anything other than my original answer. Or, to put it another way, the undoubted facts support my answer in the strongest possible way—they entail my answer—, and this kind of connection cannot be affected by adding more evidence. Moreover, I even correctly see the entailment, and initially believe my answer in virtue of seeing the entailment.

How can this be reconciled with the fact that the rationality of my confident belief in my conclusion *is* undermined by the information about the drug? It seems to me that the answer comes to something like this: In accounting for the HOE about the drug, I must in some sense, and to at least some extent, *put aside* or *bracket* my original reasons for my answer. In a sense, I am barred from giving a certain part of my evidence its due.

After all, if I could give all my evidence its due, it would be rational for me to be extremely confident of my answer, even knowing that I'd been drugged. In fact, it seems that I would even have to be rational in having high confidence that I was immune to the drug: By assumption, the drug will very likely cause me to reach the wrong answer to the puzzle if I'm susceptible to it, and I'm highly confident that my answer is correct. Yet it seems intuitively that it would be highly irrational for me to be confident in this case that I was one of the lucky immune ones. We might imagine that in the videotapes of the other subjects that I'm shown, I've seen many susceptible subjects hotly insisting that they must be immune to the drug while insisting on the correctness of

their drug-addled conclusions, which, they claim, they can just *see* to be correct. Although there's a way in which I'd be unlike them if I insisted on the correctness of my conclusion (since my conclusion is, in fact, correct), it is intuitively absurd to think that I'd be rational to argue confidently in this way in my present context. Thus it seems to me that although I have conclusive evidence for the correctness of my answer, I must (at least to some extent) bracket the reasons this evidence provides, if I am to react reasonably to the evidence that I've been drugged.

The picture of constraints on my reasoning that I'm proposing here is similar to one that some have put forth in discussing the rational response to peer disagreement. In explaining why one's confidence in the result of one's mental restaurant-check-division should drop dramatically upon learning of the friend's disagreement in the case above, I have argued (2007a) that one's assessment of the epistemic credentials of one's friend's expressed belief must be *independent* of one's own reasoning on the disputed matter. Similarly, Adam Elga (2007) holds that when I find out that my friend disagrees with me, I should assess the probability that I'm right by utilizing the *prior* probability that my answer would be right in a case of disagreement—where by "prior," Elga means epistemically prior to (and thus independent of) my own reasoning about the matter in question.[9] This sort of independence requirement would explain why—even if I happen to be the one who figured the bill correctly this time—it's not kosher for me to reason as follows: "She got $45, but the right answer is $43, so she must have made the mistake this time, and I needn't worry about her dissent."[10]

It's worth pausing for a moment to look at the bad reasoning that would be involved in these cases if agents were simply to use their first-order reasons, at undiminished strength, to support claims of their own cognitive reliability. Such reasoning would seem to beg the question in an intuitive sense, but it would not beg the question in some familiar

[9] Elga explains this notion as follows (2007, 489): "Sometimes we may sensibly ask what a given agent believes, *bracketing* or *factoring off* or *setting aside* certain considerations."

[10] See Kornblith (forthcoming) and Frances (forthcoming) for endorsements of similar principles. Some have questioned such strong independence principles (see Kelly (2005), Kelly (forthcoming), Lackey (forthcoming, a, b), Sosa (forthcoming). They hold that I may rely, *at least to some extent*, on my own reasoning to demote my peer's opinion. But it's important to see that denying that one's assessment of a peer's opinion should be *fully* independent of one's own reasoning on the disputed topic is entirely compatible with requiring that one's own reasoning on the disputed topic be discounted or put aside to at least some extent. So I'm not taking issue here with those who would deny strong independence principles of the sort I have advocated elsewhere.

ways. In classically circular arguments, one depends on a premise whose plausibility depends on one's conclusion. But in the drug case, for example, I would not be basing the conclusion that I'm immune on a premise that presupposes my immunity. My reasoning would rely only on premises provided by the puzzle, on knowing which conclusion I in fact reached, and on uncontested assumptions about the workings of the drug. So my reasoning doesn't seem to be classically circular. In rule-circular arguments, one employs a rule of inference which, if followed, would give one reliable beliefs only if one's conclusion were true. But in the bad argument for my immunity to the drug, the reliability of the inference-rules I would employ is entirely independent of my sensitivity or immunity to a certain drug. So the bad argument doesn't seem to be rule-circular either. The point of this is not that the question-begging aspect of the bad arguments in HOE cases is mysterious. But I do take it as a mark of the distinctness of HOE that it seems to require some provisions (over and above, e.g., forbidding classically circular and rule-circular justifications) to deal with this distinct sort of question-begging.

Does this requirement to bracket some of one's reasons apply only to HOE that targets deductive reasoning? I think not, though the conclusive nature of deductive reasons makes the point particularly clear. Consider the following example. I'm on the jury in a murder case, and have a great deal of evidence, including information on Jocko's whereabouts, habits, motives, and vehicle, as well as clear video-camera footage that appears to show Jocko bragging about doing the deed. The evidence against Jocko is highly compelling, and I become extremely confident, using some form of IBE, that Jocko is the killer. Then I'm given strong evidence that this morning, I was slipped an extremely powerful explanation-assessment-distorting drug.... Appropriately, I become much less confident in my hypothesis about the murderer.

It seems to me that what allows me to reduce my confidence must involve some bracketing of my IBE-based reasons for confidence in Jocko's guilt. Those reasons, though not completely conclusive, are very weighty. The first-order evidence is not in question, and the explanatory connections between that evidence and the hypothesis that Jocko is the killer remain incredibly strong. These connections, after all, do not depend on any claims about me, and the new information I learn about myself does not break these connections. I am still in possession of extremely powerful evidence of Jocko's guilt—it's just that, in this particular situation, I cannot rationally give this evidence its due, because I cannot rationally trust myself to do so correctly.

Let us compare this sort of case to a case of ordinary undercutting defeat. Suppose that, instead of learning undermining evidence about

myself being drugged, I'm given a more standard undercutting defeater: say, that Jocko had changed his appearance before the murder, and also has an identical twin brother who shares all the characteristics that seemed to tie Jocko to the crime. Surely my confidence should be undermined here as well, even though my original evidence justified extremely high confidence. And in order to take seriously the possibility that Jocko's brother is the killer, I must give up my belief in Jocko's guilt, and refrain from making the inference that originally convinced me of it.

Nevertheless, it seems to me that this second case is very different from the one in which I learn I've been drugged. In the second case, my reason for giving up my former belief does not flow from any evidence that my former belief was rationally defective. And insofar as I lack reason to worry about my epistemic malfunction, I may still use IBE, the form of inference behind my original belief, whole-heartedly. It's just that with my present, enlarged pool of evidence, IBE no longer supports the Jocko hypothesis. So the undercutting evidence does not prevent me from giving all of my evidence its due. And the fact that I must give up beliefs I formerly held does not imply that I have had to bracket any of my reasons.

Thus it seems to me that rational accommodation of HOE can require a certain kind of bracketing of some of one's reasons, in a way that does not seem to occur in accommodating ordinary evidence, even when that evidence is an ordinary undercutting defeater.

This way of looking at HOE is, I think, closely connected with the peculiarity we saw in connection with Field's work on apriority. There, it seemed that HOE sometimes required agents to violate or compromise certain rational ideals. It now seems clear why this is so. HOE can put agents in a position where they cannot trust their appreciation of the first-order reasons. So even if they see clearly what a certain epistemic ideal—such as respecting logic or IBE—requires, they cannot simply allow their beliefs to obey these requirements. That is what is meant by saying they must (to at least some extent) bracket their first-order reasons. And the result of this, of course, is that their beliefs end up falling short of the relevant ideals.

The bracketing point also connects, in an obvious way, to the person-relativity we saw above. As we've seen, HOE, unlike ordinary undercutting evidence, may leave intact the connections between the evidence and conclusion. It's just that the agent in question is placed in a position where she can't trust her own appreciation of those connections. But a different agent, who also sees the evidential relations for what they are, may (insofar as she lacks reason for self-distrust) give these reasons their full due–even though she also is

aware of the HOE that undermined the rationality of the first agent's belief.

4. HOE and Belief Revision

Another angle on the peculiarity of HOE is provided by reflection on rational revision of beliefs. Let us focus on a paradigm example of rational belief-revision, and see what happens when HOE is added to the mix.

Suppose first that I'm a rational scientist investigating some phenomenon experimentally, and suppose that, were I to get evidence E, it would give me excellent reason for confidence in H (say, because E is highly unexpected, and H would be a terrific explanation for E). And suppose it's Sunday, and I'll get the results of my experiment when I get to the lab Monday morning. In this case, it seems that these things may well be true of me:

i. I'm not highly confident that H is true.

ii. I am highly confident that if I will learn E tomorrow, H is true.

iii. If I get to the lab and learn that E is true, I should become highly confident that H is true.[11]

So the confidence in H that I should adopt Monday at the lab, if I do learn E, lines up with the confidence I have Sunday that H is true on the supposition that I will learn E on Monday.

Now instead of considering just the possible experimental outcome E, let's consider a more complex bit of evidence I could acquire Monday morning. I could learn not only E, but D: that a powerful explanation-assessment-disrupting drug is slipped into my breakfast coffee on Monday. Here, it seems that a gap opens up between the two things that lined up nicely before. First consider how confident I should be that if I will learn (E&D) tomorrow, H is true. My being drugged tomorrow has no bearing on the actual evidential/explanatory connection between E and H, and no independent relevance to H. So it seems that, today, I should think that, if E is true, H is very likely true, whether or not I get drugged tomorrow morning. Thus:

[11] Both ii and iii (and similar claims below) depend on a *ceteris paribus* assumption to the effect that I haven't gained (or lost) any other evidence relevant to H. More on this below.

iv. I am highly confident that if I will learn (E&D) tomorrow, H is true.

But if I actually do learn (E&D) tomorrow, will it in fact be rational for me to become highly confident in H? It seems not–after all, if I learn D tomorrow, it will not be rational for me to trust my assessments of explanatory support. And this is true whether or not I'm actually affected by the drug. So it seems, at least at first blush, that:

v. If I go to the lab and learn (E&D), I should *not* become highly confident that H is true.

So it seems that the HOE about my being drugged produces a mismatch between my current confidence that H is true on the supposition that I will learn certain facts, and the confidence in H that I should adopt if I actually learn those facts.

Now it might be objected that even if I learn (E & D) tomorrow, I might be able to assuage the worries induced by learning that I'd been drugged. Suppose that for some reason I could perfectly remember all of my pre-drug credences, and suppose that I also had absolute confidence in my memory, and suppose that this confidence were rational for me to have. If all that were true, it might be sufficient to defeat the defeater provided by D: the doubts about my post-drug explanatory judgments could be rationally dispelled by my being certain that my post-drug judgments were consonant with the judgments I made pre-drug. Of course, real agents don't have memories of all their previous credences, and aren't absolutely confident of those memories they do have, and wouldn't be rational in having absolute confidence in those memories even if they had it. So this sort of response would not be available to me, or to any other real person. But it might be insisted that the simple story I told above is compatible with holding that there is a match between Sunday's confidence in H supposing that I'll learn (E&D), and Monday's confidence in H upon learning (E&D), at least for an ideal agent.

I think, though, that a modification of the story would reinstate the mismatch, even for ideal agents. We might add to the description of the drug that it also distorts one's memories of one's previous credences. In that case, the potential defeater-defeater would itself be defeated: it does not seem that it would, in this revised case, be rational for an agent to be confident in H, even if she had vivid memories of her previous credences. (It's worth emphasizing that no actual interference with memory or explanatory assessments, at any time, is required

by this example. It's the *evidence* of possible interference that does all the work.) So it seems to me that insofar as HOE produces mismatches between an agent's current credence in a hypothesis on the supposition that she'll learn certain evidential propositions, and the credence the agent should adopt when she learns that those evidential propositions are in fact true, those mismatches will occur for ideal agents as well as ordinary ones.

However, there is another complication here which may serve to soften the contrast I've been trying to highlight. In the case involving the drug evidence, the way my beliefs should evolve depends crucially on my knowing my temporal location. If we take D as "I'm drugged Monday at breakfast," then D will undermine my confidence in H when I get to the lab only because I'll be confident that it's Monday morning. But on Sunday, I'm obviously not confident of *that*. So in a strict sense, one might insist that (E&D) is not, after all, all the relevant evidence that I'm getting by Monday morning. And thus, it might be urged, one would not expect that the credence I should adopt on Monday at the lab would match my former credence on the supposition just that I'll learn (E&D).[12]

I think that there is a sense in which this point is surely correct. Perhaps when we fully understand how belief updating works in contexts where self-locating beliefs are important, we will see that HOE-involving cases can be accommodated in a formal account of belief updating which preserves a general matching between present credences on the supposition that E, and future credences on learning just E. But whether or not such a formal matching principle (e.g., a form of Conditionalization applied to self-locating beliefs) is in the offing, it seems to me that, intuitively, we can see a contrast between updating involving HOE and ordinary cases of updating.

Let us again look at the case of an ordinary undercutting defeater. Suppose that in our current case, E describes the readouts of an instrument in my lab. But instead of D, let us consider an ordinary bit of undercutting evidence U: that due to an electrical surge during breakfast-time Monday, my instrument is miscalibrated Monday morning. Now we have:

> vi. I am not highly confident that if I will learn (E&U) tomorrow, H is true.

---

[12] For this reason, this sort of case can't be used as a counterexample to the standard Conditionalization model of belief revision, as I had originally thought. Thanks to Juan Comesaña and Joel Pust for pointing this out. At present, it's not clear how to handle self-locating beliefs within a Conditionalization-style model.

But this aligns perfectly with:

> vii. If I go to the lab and learn (E&U), I should not become highly confident that H.

So there seems to be a contrast between HOE and ordinary undercutting evidence in belief-updating. In the ordinary undercutter case, it's still true that between Sunday and Monday, I learn a new fact about my temporal location. But in the ordinary undercutting case, this information plays no important role, even though the power surge, like the drugging in the HOE case, occurs during Monday's breakfast.

The present phenomenon is, I think, closely related to the peculiarities we've already seen associated with HOE. We saw earlier that HOE is often agent-specific: evidence of a particular agent's impairment is relevant to her beliefs in a way in which it's not relevant to the beliefs of others. Structurally, we see the same phenomenon in the belief-updating case: HOE that would be relevant to my future self's beliefs about H is not relevant to my present beliefs about H. In this respect, D differs markedly from U.

We also saw that HOE often seems to require agents to bracket certain reasons, resulting in their failing to respect the relevant epistemic relations. This comes out in the updating case as well. I can now see that, should I learn (E & D), I'll have to bracket E, and not become highly confident in H. But I can also see that in not becoming highly confident of H, I'll be failing to give E its due, and I can see that in that situation, H is actually very likely to be true! This accounts for the sense in which the beliefs it would be rational for me to form, should I learn (E & D), are not beliefs I can presently endorse, even on the supposition that I will learn (E & D).

Again, the point is not to argue against the possibility of giving a formally unified treatment to HOE-based undermining and ordinary undercutting defeat. Rather, it's to highlight substantial differences between HOE and ordinary undercutters, differences which flow from HOE's bearing on the relevant propositions about the world only via bearing on propositions about the agent's own cognitive processes.

5. Alternatives to the "Bracketing" Picture

In arguing that HOE is best understood as requiring agents to bracket some of their reasons, I've relied on a couple of examples of rational ideals or principles that get violated by agents responding properly to

HOE. I've assumed that claims can be supported via deductive logic and via some inductive principle such as IBE. But there's a natural line of resistance to the account I've been pushing. Suppose we grant that HOE sometimes requires agents to form beliefs that violate certain deductive or inductive principles. Why see these cases as showing that agents must violate rational ideals, rather than as showing that the deductive or inductive principles in question were not rational ideals to begin with?

Here is a particularly simple way of pressing this question: Suppose we specify, for every possible evidential situation in which an agent may find herself, what the appropriate doxastic response is. The result would be an overarching rule which took into account every sort of evidence. We might then think of that rule as encoding the one and only true epistemic principle—one which, by construction, agents would never have (epistemic) reason to contravene. One might acknowledge that following this Über-rule would in certain cases lead agents to form beliefs that contravened deductive logic, or to reject hypotheses that best explained the evidence. But one could add that such an agent, by believing in accordance with the Über-rule, would *by definition* give all the evidence its due—no bracketing of any kind required!

I have no argument that this sort of description of our epistemic principles is impossible. But I also think that the possibility of describing things this way does not really cut against the view I've been defending. Moreover, I think that this sort of description would end up obscuring important features of the epistemology of HOE. Let us take up these points in turn.

First, the fact that one can definitionally integrate a number of competing ideals into a function that outputs the best way of balancing the ideals against one another does not by itself show that the original ideals are not ideals. One might, for example hold that it was morally good to keep one's promises, and also to alleviate suffering. These ideals obviously can conflict in certain situations, and one might hold that there was a morally best way of acting in such situations. Perhaps one could define a moral Über-rule which would output the morally best response in each situation. But the mere existence of this Über-rule would hardly show that keeping promises was not really a moral ideal. It might still be the case that what was morally valuable about, say, my spending time talking to my friend's plants while she was away was just that this would keep my promise to her. Keeping promises is not the only moral good, and may be trumped by other considerations. But that doesn't show that it's not intrinsically a source of moral value. Similarly, the mere possibility of defining an epistemic Über-rule would

not show that respecting logic or IBE were not epistemic ideals. Those ideals still could be what explained the reasonableness of beliefs in that were accord with them.

Second, it seems to me that the proposed way of describing our epistemic ideals would obscure an interesting point about the special epistemic role played by higher-order considerations. We've seen that certain intuitively attractive epistemic ideals have to be compromised in particular situations. But when we've seen this occur, the evidence that requires this sort of compromise is higher-order. If this is right, then it turns out that HOE (and perhaps, more broadly, an agent's higher-order beliefs) play an important and distinctive role, a role that would be hidden if we aggregated our epistemic ideals into a single abstract Über-rule. We would miss out on seeing how an agent's beliefs about herself have a sort of broad systematic effect on epistemology that's not produced by, e.g., an agent's beliefs about trees, or numbers, or other agents.

Finally, it seems to me that we should continue to recognize a sense in which there is often something epistemically wrong with the agent's beliefs after she takes correct account of HOE. There's something epistemically regrettable about the agent's being prevented, due to her giving HOE its proper respect, from following simple logic, or from believing in the hypothesis that's far and away the best explanation for her evidence. Understanding that these sorts of imperfections remain, even in the best of all possible beliefs (given the agent's evidential situation), seems to me an interesting result.

In sum: One may, by definition, rule out my description of HOE as requiring bracketing of epistemic ideals. But one would not thereby eliminate what's distinctive about this sort of evidence. And I would submit that thinking solely in terms of the Über-rule would just make it harder to see this interesting feature of the structure of epistemic rationality.

There are, of course, other ways of assimilating HOE-based undermining to ordinary undermining by rejecting the picture of reasons I've been presupposing. I've assumed, for example, that in the Drugs case, the agent's reasons for his answer to the Ferdinand question are conclusive. One obvious way of resisting this move would be by taking a more subjective view of the reasons provided by logic.

On an extremely subjective view, the reasons that support my belief about the Ferdinand puzzle do not include the fact that its truth is guaranteed by logic. Instead, my reasons would be limited to the fact that the claim *seems* logically guaranteed to me. So when learning about the drug gives me cause for distrusting my logical seemings, my reasons are undercut in the standard way: the situation would be no

different from one in which I learn about deceptive lighting, and thus have cause to distrust my perceptual seemings. One might even hold that it's precisely the necessity of allowing clear deductive reasons to be undermined in certain cases that supports the highly subjective picture.

Having made this move about deductive logical reasons, it would seem natural to make the same move about inductive reasons. One might hold that a scientist's belief in her theory is not made rational by, e.g., objective explanatory relations holding between the theory and the evidence she has, or even by her appreciation of the relations' holding. Rather the rationality would flow from its seeming to her that her evidence was best explained by the theory in question.

Although this sort of view would apparently eliminate the contrast I've been drawing between HOE and ordinary evidence, there are strong reasons to doubt that such a subjective picture will give plausible results in general. We do not think that someone who reasons in accord with, say, the fallacy of denying the antecedent attains rational belief that way, even if the conclusion he adopts strikes him as following conclusively from his premises. Similarly, a scientist whose judgments of explanatory goodness are inaccurate (say, due to emotional attachment to a pet hypothesis) does not form rational theoretical beliefs simply because she accepts theories that strike her as best explaining the data. So extremely subjective views of justification, while they might assimilate HOE undermining to ordinary undermining, are not plausible.

There are, however, views of evidence which allow the objective logical or explanatory relations to make a difference, but also take the justificatory force of these relations to depend on some subjective factor. A natural candidate is an account on which reasons are provided by the agent's *appreciation of*, or *rational insight into,* the objective logical or explanatory relations. On this sort of view, the sort of undermining produced by HOE may seem much more similar to the undermining produced in perceptual cases by evidence of deceptive lighting. How similar they turn out in the end will depend, of course, on the details of the general view of evidential support.[13] But even granting that point, I would argue that on at least some reasonable views about evidence, HOE will need to be accommodated by the sort of bracketing picture I've described above. Moreover, to the extent that HOE is seen

[13] Of course, if one was attracted to a view which gave a unified treatment to HOE and ordinary undermining evidence in part because of its unifying power, this would itself demonstrate the importance of thinking about HOE.

as involving bracketing, it will help explain our reactions to certain sorts of examples that have generated controversy.

Consider questions about how one should react to the disagreement of apparent epistemic peers. In discussing peer disagreement cases, many arguments seem to turn on a conflict between two lines of thought. In peer disagreement situations it may seem to me that when I think about my friend's arguments, I can just see how she's going wrong. But I might also know that she is generally just as smart, good at arguing, acquainted with the literature and so on as I am, and she thinks that she can see how I'm going wrong (see Kelly (forthcoming, § 5.2)). Now on the one hand, maintaining my belief in the face of this sort of disagreement can seem dogmatic. Clearly at least one of us has misjudged the arguments, and it can seem as if I'd be begging the question of who misjudged if I concluded—on the basis of the very reasoning behind my own belief on the disputed matter—that my friend made the mistake in this case. On the other hand, adjusting my belief to take account of my friend's dissent can feel irrational, in a way that it does not feel irrational at all to reduce my confidence that a table is red when I'm told that the lighting is deceptive. In both the disagreement and the lighting cases, we have undermining; but the disagreement case requires the agent to put aside what still strike her as (and what may actually be) clear reasons for her view. To the extent that our general theory of evidence casts HOE in terms of bracketing, we can explain and, in a sense, sympathize with this feeling of tension. For an agent confronted by such a situation, neither epistemic path will seem fully satisfactory.

The bracketing picture also explains why disagreement cases involving equally-good-but-distinct evidence do not elicit the same sort of vigorously mixed reactions (see Kelly (forthcoming)). Consider a case where you're a highly skilled scientist studying effects of a new drug on a sample population. Your colleague, whom you believe to be just as good an experimenter as you are, is studying the same drug on a different, but equally big, sample population. On the basis of your study, you become highly confident that the new drug is somewhat more effective than the standard therapy. Then, much to your surprise, you learn that your colleague has become highly confident on the basis of her study that the new drug is somewhat less effective. Clearly, you should become much less confident that the new drug is more effective.

This second kind of case need involve no bracketing. The information about your colleague gives you information about additional first-order evidence relevant to the issue at hand, but it does not give you strong reason to believe you've made an epistemic mistake. So becom-

ing less confident in your belief does not require you to step back from the belief that seems to you to be best supported by the (first-order) evidence. And I would suggest that this is related to the fact that no one is tempted to say that you should maintain your confidence undiminished in this sort of case.

My point here is not to plump for one side of the disagreement debate over the other. It applies even if, as some would recommend in many cases of the first sort, one should become somewhat less confident, but stop well short of splitting the credal difference with one's friend. Even that sort of moderately concessive response involves some measure of bracketing of one's reasons. And insofar as this can entail backing off, at least to some extent, of the belief that's best supported by the (first-order) evidence, it evokes discomfort with the resulting epistemic position.

6. Anti-Expertise Paradox Cases

For one more perspective on HOE's implications, let us consider the Anti-Expertise Paradox case mentioned briefly at the outset: I'm given evidence that, no matter what I do,

(B) I'll believe that P iff not-P.

A standard example involves futuristic brain-scanning devices. P might stand for "My brain is in state S," and we might imagine that thirtieth-century neuroscience, having developed accurate detectors of belief-states and of state S, has discovered that (B) very reliably holds (see Conee (1982, 57)).

A less futuristic example is mentioned in Egan and Elga (2005). Agent AE is informed that he is reliably wrong when he decides which way to turn in difficult navigational situations. It's not just his initial inclinations which are bad—even when he tries to compensate, and chooses against them, he ends up with the wrong belief about which way he should go. Egan and Elga make the sensible recommendation that AE suspend belief in the relevant situations. But they note that a variant on the example, where AE also knows that whenever he ends up suspending belief, the correct direction is left, leaves AE with no rationally stable attitude to adopt. (This more difficult version of the navigation case would substitute "I should turn left" in for P in (B) above.)

The problem in these cases is how to react to this evidence. Suppose that I believe that (B). Should I then believe P? It would seem not, since (B) tells me that in coming to believe P, I'd be adopting a false

belief. A similar problem applies to believing not-P. Should I then withhold belief about P? Even that seems bad, since given (B), I'd know in advance that I'd be withholding belief in a truth.

Paradox cases are somewhat different from the undermining cases we've been looking at. But there are striking similarities. The evidence in question is evidence of the agent's epistemic malfunction. And as in the other cases we've been looking at, the evidence is agent-specific: there is no problem with a third party rationally assessing all the relevant propositions. And it seems to me that a further similarity emerges when we think about some of the solutions that have been offered to the problem the cases pose.

Earl Conee (1987) suggests that I may withhold belief about P (and know I'm withholding), and also believe (B). But he says that I should refuse to take the obvious inferential step to believing P. So if I'm rational, I'll end up believing:

(a) I don't believe that P

and

(b) If I don't believe that P, then P,

but refusing to believe

(c) P.

A related solution is offered by Reed Richter (1990), who argues that an ideal agent would escape irrationality by believing a certain universal generalization but refusing to believe an obvious instantiation of it.

This sort of position obviously invites worries about the rationality of withholding belief on something when one has extremely strong, clear evidence for its truth. (Kroon (1983, 159) criticizes Conee's solution as depending on a "low" conception of rationality; Sorensen (1987, 312) seems to concur.) Conee acknowledges the naturalness of the worry, but in return, he replies that it can't be rational for a person in the envisaged situation to accept P (or to accept not-P), since he'd know in advance that in doing do, he'd be accepting something which would be evidently false. Conee argues that this is worse than merely failing to believe something when it is evident to one that it is true. He concludes that his suggested resolution is the option with "the highest epistemic value" (1987, 326).

A very different approach is suggested by Roy Sorensen (1987). On his view, the rational course is to refuse to believe (B). This response obviously invites worries over whether refusing to believe (B) in the face of what might be massive evidence can really be rational. In the brain-device case, one can imagine that the device has been tested extremely thoroughly, perhaps on 999 exact clones of the agent (see Richter (1990, 150–151)). In the navigation case, AE may have an incredibly long track-record of bad judgments he's made, many of them made in full realization of long track-record support for (B).

Sorensen argues that no conceivable evidence could be strong enough to make belief in (B) rational. Since believing (B) would necessitate violating conditions of rationality (which require, for example, knowledge of one's own beliefs, and following through on the Modus Ponens from which Conee would have the relevant agents abstain), an ideally rational agent cannot accept (B) on *any* evidence. Sorensen writes: "Since we are warranted in making costly revisions to escape acceptance of an inconsistent proposition, we are also justified in paying a high price to avoid positions which cannot be consistently accepted" (1987, 312).[14]

I don't want to address here the question of which (if either) sort of response to the problem cases is the best. Instead, I want to notice how the solutions we've been considering resemble the reactions discussed above to less dramatic cases of undermining HOE. The troubling aspect of the paradox situation consists in the fact that the agent seems to have to violate some rational ideal, no matter what she does. Different writers find different solutions superior. Conee argues that it's rational for the agent to withhold on a proposition even though she has excellent evidence for its truth, on the grounds that the other options are worse. But the criticism that Conee's solution employs a "low" conception of rationality may be seen as a giving voice to a worry that would apply even if Conee is right that withholding has the highest epistemic value: that there is something epistemically regrettable in the option Conee recommends. Similarly, Sorensen's comment that we are justified in "paying a high price" for avoiding inconsistency seems to reflect recognition that refusing to believe (B) in the face of the evidence exacts an epistemic price. This is the problem that bothers critics of Sorensen's position (e.g. Richter (1990)), but it seems to me that it would apply even if Sorensen has correctly identified the most rational response to the cases.

---

[14] Sorensen also offers more specific reasons for doubting that the sort of empirical evidence one might have for (B) should be persuasive. Sorensen's arguments are discussed in Richter (1990).

In the cases considered earlier, I argued that the epistemically best response an agent could give would still fall short of some rational ideal.[15] The difference is that in many ordinary cases of HOE, it seems clearer what a person should do. Perhaps this difference is due to the fact that the violation of rational ideals in those cases is less extreme, or more familiar, so it's easier to see the end result as unproblematic, or at least non-paradoxical. But it's also worth noting that people do differ strongly on the right reaction to certain ordinary HOE undermining cases, as is evident from the literature on peer disagreement. And in conversation, I've encountered a wide spectrum of opinions on how much my confidence in my answer to the Ferdinand puzzle should be shaken in the Drugs case. Some see it as obvious that if the drug causes 80% of subjects to become confident in wrong answers, I should be no more than 20% confident in my answer (even if I happen to be immune to the drug). Others feel equally strongly that in this situation, I should remain fully confident of my answer. And still others come down somewhere in between. In a way, then, the anti-expertise paradox cases represent the far end of a spectrum of epistemic difficulty engendered by taking account of HOE.

This perspective on the paradoxical cases seems to me also to shed light on cases where evidence of anti-expertise is less virulent, such as the majority of cases analyzed by Egan and Elga. They concentrate on cases where the agent gets strong evidence that she's highly unreliable about a certain range of matters, but the agent can withhold belief about the matters in question, and has no evidence suggesting that her withholding strongly indicates the truth or falsity of the matters in question. Their purpose is not to explore paradox; in fact, they see the cases they study as admitting of perfectly reasonable solutions. Their purpose is to show that there are clear limits on the degree of anti-expertise an agent can attribute to herself while remaining coherent and having reasonably good access to her own beliefs, conditions they take as constituents of rationality. Interestingly, their conclusion in these cases parallels Sorensen's conclusion about the paradoxical cases: they hold that one can't rationally self-attribute certain sorts of anti-expertise, a result they summarize as "I can't believe I'm stupid".

Egan and Elga mean, of course, that I can't *reasonably* believe I'm stupid. But from the perspective defended above, this conclusion must be treated carefully. Suppose we take it that in the paradoxical cases, one is forced to violate some ideal of rationality or other. So there's a

[15] Frederick Kroon (1983) expresses a similar view of the paradoxical cases, though he later (1993) rejects it.

sense in which the evidence of anti-expertise compromises one's rationality. That doesn't yet peg any particular belief as irrational. So we must in general distinguish between the following two claims, where P is some self-attribution of anti-expertise:

(1) Believing P would entail my violating some rational ideal.

(2) Believing P would be irrational in the sense that taking some other attitude toward P would be rationally superior.

Applying this distinction to non-paradoxical cases, we might grant Egan and Elga's central result, which involves a claim of type (1): that in these cases, an agent who self-attributes anti-expertise ends up violating some rational ideal. But we might still wonder whether, in all such cases, the irrationality entailed by self-attribution of anti-expertise is, on the model of (2), *located in* that self-attribution.

Egan and Elga recommend that agents escape irrationality by giving up the beliefs that are the target of the anti-expertise evidence; having given up the target beliefs, an agent no longer has reason to consider herself an anti-expert. But Nicolas Bommarito (2009) provides a kind of example that casts doubt on whether, even in some variants of the non-paradoxical sorts of cases considered by Egan and Elga, we should see the self-attribution of anti-expertise as itself irrational. Consider the following example:[16]

Suppose that I encounter very strong evidence that I'm inescapably drawn to some inaccurate way of thinking. Perhaps a very credible psychologist tells me that I underestimate the intelligence or competence of people with certain accents, or skin colors. There's a sense in which, ideally, I should simply compensate for my bias, but the psychologist might well tell me that any attempts I make at compensation will fail. Can I then simply eliminate the problematic beliefs by withholding—by eliminating my beliefs about certain people's intelligence or competence? That, too, might be desirable, but I may well get evidence that it's not psychologically possible for me simply to abstain from such judgments about people. In this sort of situation, it is far from obvious that self-attributing anti-expertise is itself irrational, even if doing so guarantees that I'll end up in violation of certain ideals of coherence. It could be that, given the psychological facts of life, "believing I'm stupid" is the only rational attitude to take toward myself.

[16] This example is not Bommarito's, but its essential feature, what he calls "sticky irrationality," is taken from one of his examples.

Note that the point here doesn't depend on the claim that, because I'm in fact a limited agent, I'm vulnerable to prejudice, and I'm psychologically unable to shake the prejudiced beliefs. The argument would go through even if I were fully able to control my beliefs, and even if I weren't at all subject to prejudice. After all, even an agent who did not have either of these limitations could get *evidence* that she did have them. And it's the (higher-order) evidence that seems to demand the self-attribution of anti-expertise.

This conclusion parallels the results we saw in other cases of HOE. If I get powerful evidence that I've been drugged in a way that will prevent me from trusting my appreciation of logical or explanatory support relations, respecting that evidence may require my violating certain rational ideals.[17] But this doesn't show that respecting that evidence would be irrational in sense (2). In fact, I've argued that in some such cases, respecting the HOE is rationally required, despite the evident costs of doing so.

7. Conclusion

The position I've been defending might be put metaphorically by saying that one of the peculiarities of HOE seems to be that it's prone to being rationally toxic: that is, being such that once the agent has it, she is doomed to fall short of some rational ideal. Of course, HOE may come in degrees of toxicity, with the paradox cases involving severely toxic HOE.

On this picture, it's important to see that toxic evidence situations need not be situations in which there's no rationally best response. It may be that our epistemic ideals are not all simultaneously realizable in certain situations; but that does not show that any old way of partially realizing them is equally good, or even that there isn't some most rational way.

One might, of course, count the most rational way of reacting to a certain situation as, by definition, rationally perfect. That would be essentially to adopt the picture of rationality as being governed by one Über-ideal, that of achieving the best balance among the prima facie ideals that HOE requires agents to violate. On this picture, there would in a sense be no such thing as toxic evidence, by definition. But, as argued above, this picture would hide interesting aspects of the way HOE plays a distinctive role in our rational lives.

Insofar as we do see HOE as liable to toxicity, it raises the question of why, if HOE has such toxic effects, we are in general rationally required to respect it—don't we have enough trouble living up to rational ideals as it is?

---

[17] I take the phrase "respecting the evidence" from Feldman (2005).

The answer, it seems to me, flows precisely from something implicit in the question: that we unfortunately have no guarantee that we're living up to the ideals of rationality. I have been concentrating, in order to highlight HOE's distinctness, on cases where one has managed to reach one's initial belief completely rationally, and then HOE undermines one's judgment. In those cases, it is indeed true that one's epistemic health is compromised by HOE. But equally, in those all-too-common cases when one has made an initial error epistemically, the warning provided by HOE can serve to ameliorate the malady.

In fact, evidence of our epistemic frailty is pervasive, and taking appropriate precautions is an essential part of any rational human life. The medical resident may thus be saved from acting on his fatigue-induced misdiagnosis. The same applies to the judge who recuses herself from a case she knows she's emotionally invested in, because she knows that what seems just to her may be distorted by her attachments. (And, some would argue, the self-confident philosopher may be saved—not so dramatically or practically, but intellectually—by backing off of his confidence in mistaken views, when he learns that equally competent philosophers, familiar with all the same arguments, see things differently).

In fact, as these examples suggest, it is often an epistemically sound strategy not only to respect HOE, but to seek it out. This is what goes on when the engineer gets a colleague to check the specifications of her bridge design, or the doctor recommends to his patient to seek a second opinion. HOE is, it seems, an extremely valuable resource for creatures living with the possibility of imperfection. It is an indispensible epistemic tool, not an isolated curiosity.

The present essay is clearly just a first step toward understanding this sort of evidence. At a minimum, even if the picture I've been supporting is basically correct, the notion of bracketing surely needs further explanation and development. I'm not sure how much of this can be done independently of particular overall accounts of epistemic rationality; after all, as noted above, the necessity of invoking something like bracketing at all will depend on the overall shape of one's theory of rational belief. But it seems to me that any satisfactory epistemology will need to address the apparent peculiarities with which HOE presents us. We will not fully understand rational belief until we understand higher-order evidence.

References

Bommarito, N. (2009) ''Rationally Self-Ascribed Anti-Expertise,'' *Philosophical Studies* (published online first: DOI 10.1007/s11098-009-9441-3).

Christensen, D., (2007a), ''Epistemology of Disagreement: The Good News,''. *Philosophical Review* 116: 187–217.
—— (2007b), ''Does Murphy's Law Apply in Epistemology? Self-Doubt and Rational Ideals,'' *Oxford Studies in Epistemology* 2: 3–31.
Conee, E. (1982), ''Utilitarianism and Rationality,'' *Analysis*, Vol. 42, No. 1: pp. 55–59
—— (1987), ''Evident, but Rationally Unacceptable,'' *Australasian Journal of Philosophy* 65: 316–326.
Egan, A and Elga, A. (2005), ''I Can't Believe I'm Stupid,'' *Philosophical Perspectives* 19: 77–93.
Elga, A. (2007), ''Reflection and Disagreement,'' *Noûs* 41: 478–502.
Feldman, R. (2005), ''Respecting the Evidence,'' *Philosophical Perspectives* 19: 95–119.
—— (2006), ''Epistemological Puzzles about Disagreement,'' in S. Hetherington (ed.) *Epistemology Futures* (New York: Oxford University Press).
Feldman, R. and Warfield, T. eds. (forthcoming), *Disagreement* (Oxford: Oxford University Press).
Field, H. (2000), ''Apriority as an Evaluative Notion,'' in P. Boghossian and C. Peacocke, eds., *New Essays on the A Priori* (New York: Oxford).
—— (2005), ''Recent Debates about the A Priori,'' in T. Gendler and J. Hawthorne, eds., *Oxford Studies in Epistemology* (Oxford University Press): 69–88.
Frances, B. (forthcoming), ''The Reflective Epistemic Renegade,'' *Philosophy and Phenomenological Research*.
Kelly, T. (2005), ''The Epistemic Significance of Disagreement,'' *Oxford Studies in Epistemology* 1: 167–196.
——, (forthcoming), ''Peer Disagreement and Higher-Order Evidence,'' in Feldman and Warfield.
Kornblith, H. (forthcoming), ''Belief in the Face of Controversy,'' in Feldman and Warfield.
Kroon, F. (1983), ''Rationality and Paradox,'' *Analysis* 43: 455–461.
—— (1993), ''Rationality and Epistemic Paradox,'' *Synthese* 94: 377–408.
Lackey, J. (forthcoming a), ''A Justificationist View of Disagreement's Epistemic Significance,'' in A. Haddock, A. Millar and D. Pritchard, eds., *Social Epistemology* (Oxford: Oxford University Press).
—— (forthcoming b), ''What Should We Do when We Disagree?'' in T. S. Gendler and J. Hawthorne, eds., *Oxford Studies in Epistemology III* (Oxford: Oxford University Press).
Pollock, J. L. (1986), *Contemporary Theories of Knowledge* (Totowa, NJ: Roman and Littlefield).

Richter, R. (1990), "Ideal Rationality and Hand-Waving," *Australasian Journal of Philosophy* 68: 147–156.
Sorensen, R. A. (1987), "Anti-Expertise, Instability, and Rational Choice," *Australasian Journal of Philosophy* 65: 301–315.
Sosa, E. (forthcoming), "The Epistemology of Disagreement," in his *Armchair Philosophy* (Princeton University Press, 2010).