

Does Murphy's Law Apply
in Epistemology?

David Christensen

[Very rough working draft, missing citations (and more).]

1. Ideal vs. Human-centric Rationality

What would an ideally rational agent believe? Of course, the answer depends on just what kind of ideally rational agent is in question. But when epistemologists consider this question, they don't simply answer "everything true". Rationality, after all, involves reacting correctly to the evidence one has, but does not seem to require having all possible evidence about everything. Thus if we seek to understand rationality by constructing a model of ideally rational belief, we will not concentrate on an omniscient being. Instead, we'll consider a non-omniscient thinker who nevertheless is in certain respects cognitively perfect. We might, for example, stipulate the following kinds of things about such an ideally rational agent's beliefs: They would not be based in wishful thinking. They would not be biased by the agent's likes and dislikes. They would not reflect mistaken logical reasoning. Let us call an agent who is ideally rational in this sense an IRA.

This general approach to theorizing about rationality dovetails nicely with the tradition which relates rationality to thinking logically, and then characterizes rational belief with the aid of formal logic. After all, if an agent were ideally rational, we might think that its beliefs would respect logic completely. Those who see belief as a binary, all-or-nothing, kind of state have thus often taken logical consistency and logical closure to be rational ideals. And those who conceive of beliefs as coming in degrees have taken conditions based on probabilistic coherence--which can be seen as little more than applying standard deductive logic to graded beliefs--as an ideal.

(By "conditions based on probabilistic coherence," I mean not only conditions requiring agents to have precise real-valued degrees of confidence satisfying the laws of probability, but also conditions modeling rational degrees of belief by sets of probability functions, or qualitative probabilities. For

convenience, I'll use the term "probabilistic coherence" to refer to this whole family of conditions.)

Of course, this whole formal approach to thinking about rationality has been criticized. The main line of criticism takes off from the fact that ideals such as logical consistency or probabilistic coherence are very clearly far beyond the capacities of any human to achieve--even more so than complete freedom from prejudice or wishful thinking. Such ideals require, for instance, that an agent believe (or, in the case of coherence conditions, be completely certain of) every logical truth. Why then, it is asked, should rules that might apply to a peculiar sort of imaginary beings--ideal thinkers with limited information--hold any interest us? The fact that we humans have the particular limitations we do, it is urged, is not just some trivial footnote to epistemology; it's a central aspect of our epistemic predicament. Interesting epistemology--epistemology for humans--must take account of this fact.

Now the interest of any idealization depends on the project in which it is supposed to play a part. If one's epistemological project is characterizing our ordinary casual way of using "rational" and "irrational" to apply to people--a way that counts most of us as rational--then it may be hard to see how a humanly unattainable ideal will play an important role. But there's little reason to think that epistemology should be restricted to such a thin notion of rationality. Similarly, ethics should not be restricted to studying moral ideals that are perfectly attained by the ordinary people we'd hesitate to call "immoral."

A similar point applies to the project of developing a notion of rationality that's closely linked to an "ought"-implies-"can" notion of epistemic responsibility. Clearly, we don't want to blame anyone for failing to live up to an unattainable ideal. But there are certainly evaluative notions that are not subject to "ought"-implies-"can". I would argue that our ordinary notion of rationality is one of them: when we call a paranoid schizophrenic "irrational", we in no sense imply that he has the ability to do better.

Another epistemic enterprise in which the importance of highly idealized models might be questioned is the so-called "meliorative project"--epistemology aimed at our cognitive improvement. Some have claimed that any interesting

epistemology must be aimed at providing us with guidance to help ourselves (or perhaps others) to think better. I personally doubt that philosophers are particularly well-equipped for this sort of endeavor. But even putting that doubt aside, I see no reason to think that the sole point of epistemology should be the production of manuals for cognitive self-help.

What projects are there, then, which make manifestly unattainable epistemic ideals worth studying? One such project is that of assessing us as a species. After all, there is no reason to suppose--even if we are the cognitive cream of the mammalian crop--that we're the be-all and end-all of any evaluative epistemic notion we come up with. Indexing epistemic perfection to the cognitive capacities of homo sapiens clearly begs some interesting questions.

But the most important reason for resisting the impatience some express about idealized models of rationality does not depend on the interest of evaluating humans as a species. It is clear that our ordinary rationality judgments are based in assessments of peoples' levels of performance along certain dimensions of epistemic functioning. And these dimensions may well be ones whose extremes are beyond human reach. Freedom from wishful thinking is a plausible example. Predicting consequences of social policies in a way that's untainted by self-interest is another. More examples include evaluating other people's behavior and character without prejudice from emotional ties, or from bigotry based on race or sexual orientation. And a natural candidate for this list is reasoning in a way that's free of logical error.

If rationality consists (at least partly) in good performance along this sort of dimension, then one natural approach to understanding rationality more clearly is to study candidates for rationality-making qualities by abstracting away from human cognitive limitations, and considering idealized agents who can perfectly exemplify the qualities under consideration. Is logical consistency of all-or-nothing belief a rational desideratum? What about probabilistic coherence of degrees of confidence? How should agents update their beliefs when presented with new evidence? All of these questions may be approached, at least in part, by asking ourselves, "What would an IRA believe?"

Now it is important to see that the suggestion here is not that questions about rational ideals reduce to questions about what ideal agents would believe. Any such reduction would run afoul of immediate counterexamples involving, e.g., beliefs about the existence of ideally rational agents. For example, it might be the case that any ideally rational agent would be quite confident that there were highly intelligent beings who could not remember making any cognitive errors; this does nothing to show that such a belief is rationally mandatory in general. But this sort of problem does not, I think, undermine the usefulness of IRAs in studying rationality. It's just that one has to be alert to the distinction between those aspects of an IRAs beliefs which help make it ideally rational, and those that are mere side-effects of the idealization.

It might be insisted that we must still connect our thoughts about IRAs with claims about us non-ideal agents. However, there are simple, plausible ways of doing this. For example, one attractive thought is that if the constraints that apply to IRAs describe the endpoint of a spectrum, then the closer an actual agent's beliefs are to that end of the spectrum, the better (presumably, *ceteris paribus*). Efficiency in cars is a nice analogue here: perfect efficiency is impossible, but (*ceteris paribus*) the more closely one approaches this end, the better. Moral principles also might work this way: I am undoubtedly psychologically incapable of being perfectly fair or generous; but the more closely I approximate perfect fairness and generosity, *ceteris paribus*, the better.

To my mind, some of the most promising applications of highly idealized theorizing about rationality involve taking probabilistic coherence as a constraint on degrees of belief. The considerations rehearsed above, I think, show that some of the most common objections to idealizations involving probabilistic coherence, on the grounds that it abstracts so far from human limitations, are misguided. I would like, then, to say something like: "Well, of course none of us can be probabilistically coherent, but that's no big deal. We can see that coherence is an ideal by showing that IRAs have coherent credences. And as far as my own beliefs are concerned, the closer I can come to having coherent credences, the more rational my beliefs will be."

Unfortunately, I now think that this claim about IRAs just isn't true; and I think that the claim about my own beliefs is at least problematic. The reasons for this are quite different from the worries about idealization I just described. They raise what seems to me an interestingly different difficulty for the standard way of using ideal agents in theorizing about rationality, a difficulty flowing from the structure of our epistemic ideals.

2. Ideal Rationality Meets Possible Cognitive Imperfection

The problem I would like to examine involves a very different way in which cognitive imperfection poses an obstacle to taking probabilistic coherence as a rational ideal. The problem arises from an agent's apparently rational reflection on its own beliefs. Let us begin by thinking about a case involving an ordinary, non-ideal agent:

Suppose I prove a somewhat complex theorem of logic. I've checked the proof several times, and I'm extremely confident about it. Still, it might seem quite reasonable for me to be somewhat less than 100% confident. I should not, for example, bet my children's lives against a cup of coffee that the proof is correct. After all, balancing my checkbook has shown me quite clearly that my going over a demonstrative argument repeatedly is not sure proof against error. Given my thorough checking, my being in error this time may be incredibly unlikely; nevertheless, it is hard to deny that it has some nonzero probability. Let us call the theorem I've proved T. And let us call the proposition that, in "proving" T, I've "proved" a false claim by mistake M. The question now arises: given this sort of doubt, how strongly--ideally speaking--should I believe T?

It seems that my giving some slight credence to M is required by my recognition that I may sometimes exhibit cognitive imperfection. And to the extent that I have any rational credence at all in M, I must have some rational credence in the negation of T (since M obviously entails $\sim T$). So my confidence in T should fall short of absolute certainty; in probabilistic terms, it should be less than 1.

But if something like this is correct, it seems to raise an obstacle to taking coherence as a rational ideal for me--an

obstacle quite different from that raised by the contingent fact that coherence is humanly unattainable. For according to this argument, it would not be rational for me to have full confidence in T, a truth of logic. In fact, if I did manage to have the coherence-mandated attitude toward T, the argument would urge me to back away from it. So the problem is not the usual one cited in connection with human cognitive limitations. It's not that I can't achieve the probabilistically correct attitude toward T--in this case, I may well be perfectly capable of that. The problem is that, in the present case, it seems that my beliefs would be worse--less rational--if I were to adopt the attitude toward T that's mandated by probabilistic coherence.

Clearly, this point should be disconcerting to those of us who would advocate coherence--either the simple version, or one of the standard generalizations--as a component of ideal rationality. To my mind, the threat it poses is significantly deeper than that posed by the fact that probabilistic perfection is not humanly possible. Thus it's worthwhile seeing whether the one might resist the claim that it would be irrational for me to be certain of T.

3. Can I Rationally be Certain of T?

Suppose one were to argue as follows:

Certainty Argument: Granted, I must give $\sim T$ at least as much credence as I give to M. But I have the strongest possible kind of justification for full confidence in T--I've proved it demonstratively. So I should give it full confidence, and should give $\sim T$, and thus M, zero credence. (After all, my proof of T serves as a proof of not-M!) I may not be a perfect being, but I have the best possible reason for believing T, and thus the best possible reasons for being certain that I haven't "proved" a false "theorem" by mistake.

I think that this argument should not tempt us. To see why, suppose that I work out my proof of T after having coffee with my friend Jocko. Palms sweaty with the excitement of logical progress, I check my work several times, and decide that the proof is good. But then a trusted colleague walks in and tells me that Jocko has been surreptitiously slipping a reason-

distorting drug into people's coffee--a drug whose effects include a strong propensity to reasoning errors in 99% of those who have been dosed (1% of the population happen to be immune). He tells me that those who have been impaired do not notice any difficulties with their own reasoning--they just make mistakes; indeed, the only change most of them notice is unusually sweaty palms. Here, my reason for doubting my proof, and the truth of T, is much stronger. It seems clear that in the presence of these strong reasons for doubt, it would be highly irrational for me to maintain absolute confidence in T. Yet the certainty argument would, if sound, seem to apply equally to such extreme cases.

Could this verdict possibly be resisted? Could one argue that, initial appearances to the contrary, we actually can embrace the certainty argument, even in the strong doubt case? One way of attempting this would capitalize on distinguishing carefully between two sorts of cases: the bad ones, where the drug has impaired my reasoning and my proof is defective, and the good ones, in which I'm one of the lucky 1% who is immune to the drug's effects and my proof is correct. It might be pointed out that we cannot assume that what would be irrational for the person in the bad case would be irrational for someone in the good case. After all, those in the good case have constructed flawless sound proofs of T, and those in the bad case have made errors in reasoning. To say that what holds for one must hold for the other would be to conflate having a correct proof with seeming to oneself to have a correct proof. So it might be argued that although it would be clearly wrong for most people who find out that they've been dosed to dismiss the resulting doubts, if I am actually in the good case, I am in a different epistemic position, and I may rationally dismiss the doubts.

Now I think that there is something to this point. I would not claim that the epistemic situations of the drug-sensitive person and the immune person are fully symmetrical. After all, the drug-sensitive person makes a mistake in reasoning even before she finds out about the drug, and the drug-immune person does not. But granting the existence of an asymmetry here does not mean that it is rational for the drug-immune person to disregard the evidence suggesting that she has made an error. And it seems clear--especially when one keeps in mind that those who are affected by the drug don't notice any impairment in their reasoning--that given the evidence suggesting I've made a

mistake, it would be irrational for me to maintain full confidence in my reasoning, even if I happen to be in the good case.

Thus we cannot exploit the real epistemic asymmetry between the drug-sensitive and drug-immune people to argue that the latter may after all avail themselves of the certainty argument. And if this is correct, it is hard to see how we can support applying the Certainty Argument even to the original cases involving mild self-doubts raised by memories of misadventures in checkbook balancing. Nothing in the Certainty Argument hinged on the mildness of the doubt about my proof. Even if my reason for doubt is slight, and, so to speak, metaphysical,--so slight that in ordinary cases, I wouldn't bother to think about it--still, it would seem irrational to be absolutely certain that I have not erred. And thus it would seem irrational for me to be absolutely certain of $\sim M$, or of T.

If this is right, it underlies a troubling result for those of us who see coherence as a rational ideal. For the only way I can live up to the ideal of coherence here would seem to be by irrationally ignoring the possibility that I've made a mistake in proving T. Being certain of logical truths seems not only to be something that I can't always do--it seems like something I often shouldn't do. And that makes it hard to see what kind of an epistemic ideal probabilistic coherence could be.

4. Can an Ideally Rational Agent be certain of its own ideality?

The troubling result flows from the fact that I must believe myself to be epistemically fallible. But if rational ideals can be thought of as those whose attainment would make the ideal agent's beliefs rational, perhaps this is not the right way to think about the issue. Perhaps an ideally rational agent would not only never make a cognitive error, but would also (rationally) be certain of its own cognitive perfection. If that were so, then we could at least hold that an ideally rational agent would have probabilistically coherent beliefs. And this might help explain a sense in which coherence was, after all, an epistemic ideal. (I should note that this would not solve the whole problem. We would still need to say something about how ideals that apply to such imaginary agents would relate to rationality assessments for humans. I'll return to this issue later.)

It has been claimed that ideal intellects have this sort of self-confidence. For example, Jordan Howard Sobel argues that ideal intellects not only are probabilistically coherent, but display a number of other features as well: They are always absolutely certain, and correct, about their own credences. They have the sort of absolute confidence in their future credences that is embodied in van Fraassen's principle of Reflection. And they are absolutely certain that they are probabilistically coherent. Thus an ideal intellect would not only be absolutely certain of T--she'd also have the sort of high intellectual self-opinion that would seem to be needed to be rationally certain that $\sim M$.

Could we, then, hold that IRAs may believe in accordance with the Certainty Argument, and have full confidence that they have not, e.g., been caused by a drug to "prove" a false "theorem" by mistake? It seems to me that reflection on the motivation behind theorizing about IRAs should make us wary of such a move.

As noted above, an IRA, as usually conceived of in theorizing about rationality, is quite different from an omniscient god. The IRA reasons perfectly (and is thus logically infallible), but is not assumed to be factually omniscient. This conception of an IRA carries with it no obvious presumption that an IRA would know that it was ideally rational. The claim that it was ideally rational would seem to be an ordinary contingent proposition, belief in which would need some sort of a posteriori warrant. And although it seems likely that many IRAs would have excellent evidence of their rational prowess, it also seems unlikely that all of them (or perhaps any of them) could be rationally certain of their own rational ideality.

If an IRA has been around for a long time, and if it had a good memory, it might well have evidence that it possessed an excellent epistemic track record. Unlike most of us, it would never have been corrected for a cognitive error. But it's hard to see how even a very long and distinguished epistemic history could justify the sort of absolute self-confidence at issue here. For it's clearly possible for an agent to think flawlessly for a very long time, and then to make a mistake. Clearly, the spotless record does nothing to tell against this possibility.

It's also difficult to see how an agent could be introspectively aware of its own cognitive perfection--or, more precisely, it's hard to see how any sort of introspective awareness could justify absolute self-confidence. Anyone who has experienced some of the common states of consciousness involving diminished cognitive capacities knows that, in some cases, it's pretty easy to tell introspectively that one is epistemically impaired. But not all impairments are evident in this way (and even if they were, there's no reason to think that all possible impairments would be). So the fact that an agent seems to herself to be thinking with perfect lucidity could hardly justify absolute epistemic self-confidence.

For these reasons, it seems unlikely that an IRA would be rationally certain of its own cognitive perfection.

5. Can an Ideally Rational Agent be certain of T?

It might be objected, however, that the whole line of argument in the previous section is misdirected. After all, what's directly at issue in our example is just whether the IRA can rationally be certain of $\sim M$. The broader claim I just discussed, which concerns the agent's own general rational perfection, is clearly a contingent proposition, and as such might not be known by the IRA. But $\sim M$ follows from T--it's a truth of logic! So the fact that the IRA can't rationally be certain that it never makes logic mistakes is simply irrelevant. The IRA has solid a priori reason to be certain that in proving T, it hasn't "proved" a false sentence by mistake. No reliance on track records or introspection is required.

Although this argument rightly points out a disanalogy between $\sim M$ and general epistemic self-confidence, it seems to me that the disanalogy will not suffice for the ultimate use which the argument would make of it. We should first note in general that the fact that an agent has a priori justification for some belief does not render her justification immune to undermining or rebutting by a posteriori considerations. If it were, then that even in the case where Jocko tells me that he drugged my coffee, I would be justified in continuing to believe T. But given that even a priori justifications are vulnerable in this way, it's not clear why the IRA's justification for being absolutely confident in T isn't undermined by uncertainty about its own cognitive perfection.

We can see this point from a different angle by supposing, as the objection urges, that the IRA may be less than certain of its own logical prowess, but may nevertheless be fully certain of both T and $\sim M$. If the agent can achieve certainty about T and $\sim M$, it presumably can perform similar feats a great many times--there is nothing special about T. For all of the apparent theorems (say T1 - T10,000) the agent proves, it may rationally be certain of the corresponding propositions ($\sim M1$ - $\sim M10,000$) each denying that it had mistakenly "proved" a false sentence. Assuming that it could remember which claims it had constructed apparent proofs for, it could get as much evidence as it wanted for the contingent proposition that it was ultra-reliable in constructing proofs!

So granting that the agent is rationally certain of propositions like $\sim M$ seems to lead to the claim that the agent can, after all, rationally reach extremely positive conclusions about its own logical prowess. And although this sort of inductive reasoning would not produce certainty, it is still intuitively irrational. After all, in each case, the agent's reason for thinking its proof of Tn was not mistaken is simply the agent's proof of Tn. The agent's procedure here would be like consulting the gas gage in your car repeatedly, each time using it to determine both the level of fuel and what the gas gage read, and then concluding that the gage was accurate. Or looking at colored squares repeatedly, each time determining what color the square was and how it looked, and then concluding that your color vision was accurate. Insofar as the agent has any rational doubts about its logical prowess, it seems clear that the bootstrapping procedure described above cannot legitimately allay them.

The bootstrapping problem is just a reflection of the basic fact that lies behind all of the examples we've looked at: that the rationality of first-order beliefs cannot in general be divorced from the rationality of certain second-order beliefs that bear on the epistemic status of those first-order beliefs. This is the reason that, in the case of an ordinary logician who has proved a theorem, empirical evidence about being drugged in certain ways can undermine a belief whose justification was purely logical. Thinking about $\sim M1$ - $\sim M10,000$ is simply a way of amplifying a point that applied to the original $\sim M$: that insofar as an agent is not absolutely confident in its own logical faculties, it is likely to be irrational for it to be

absolutely confident in particular beliefs delivered by those faculties.

Does this point apply to even the most simple and obviously self-evident-seeming beliefs? If not, there may be a different way to argue that the IRA may have full confidence in $\sim M1 - \sim M10,000$. Consider a logical truth that, to us, is maximally obvious--say, an instance of the law of the excluded middle or the law of non-contradiction. For example, take

T': It's not the case both that a chair is in the room and that no chairs are in the room.

Even if it were granted that we would be irrational to place full confidence in complex logical theorems, it might be claimed that we should at least be able to be absolutely confident in claims such as T'. And if so, we ought to be able to have full rational confidence in the negation of

M': In believing that T', I'm believing a false claim due to a cognitive mistake.

But the IRA, it might well be argued, would experience all logical truths, including $T1 - T10,000$, as being just as self-evident as T'. So it would, after all, be rational for such a being to be completely certain of $\sim M1 - \sim M10,000$. And even if it used these claims as inductive support for a conclusion about its own logical prowess, that would be legitimate; the air of illegitimacy arises only if we're misled by ignoring the IRA's superior ability to see clearly and distinctly in cases where we cannot.

It seems to me that this strategy for supporting the IRA's certainty about $\sim M1 - \sim M10,000$ will not work. For even if we grant that all theorems seem to it as simple and obviously self-evident as T' seems to us, I doubt that the seeming obviousness or self-evidence of T' licences us in being absolutely certain of $\sim M'$. To my mind, T' is just a bit more obvious-seeming than some claims I now know to be false: that for any condition, there's a set of things that fit that condition, or that if one set of things is properly included in another, the members of the two sets can't be matched up 1:1. And even if there were some special sense of clarity and distinctness that I got only when contemplating claims like T', I would be hard-pressed to be

certain that no drug or demon could make me have that sense when contemplating a logical falsity.

For similar reasons, the IRA, even if all logical truths strike it the way T' strikes me, should not be absolutely confident in $\sim M1 - \sim M10,000$. And if that's right, it cannot rationally be absolutely confident in $T1 - T10,00$; which is to say that it cannot be probabilistically coherent.

If the argument of the last two sections is right, then, we are faced with the following sort of problem: given that an agent has limited evidence, it turns out that there is a tension among three prima facie appealing (if somewhat loosely formulated) rational ideals:

1. (LOGIC) Rationality requires respecting logic; in particular, ideally rational beliefs must satisfy (some version of) probabilistic coherence.
2. (EVIDENCE) Rationality requires proportioning the agent's beliefs (at least about logically contingent matters) to the agent's evidence.
3. (INTEGRATION) Rationality requires that an agent's object-level beliefs reflect the agent's meta-level beliefs about the reliability of the cognitive processes underlying its object-level beliefs.

The problem we saw was that if an ideal agent with limited evidence satisfied (1) with respect to its beliefs about logical theorems, and (2) with respect to its beliefs about its own cognitive processes, it could not respect (3) with respect to the connections between these two kinds of beliefs.

There are several different reactions possible here. One could of course take the problem as showing that there's something wrong with at least one of the purported rational ideals--at least, in the ways I've been interpreting them. Since I find each of them quite compelling, though, I'd like to explore other options. The first is to trace the problem to the standard sort of IRA I've been discussing.

6. Can Variant IRAs Avoid the Tension?

If the conflict among (1) - (3) arises only because we are taking our IRA to have incomplete evidence, might we avoid the whole problem by simply dropping this assumption? After all, God, on some standard conceptions, is an agent who is not only perfectly rational, but also perfectly informed. It can be hard to understand how God knows things--it would seem that nothing like our ordinary sources of empirical evidence would be necessary (or, really, of any use at all) for God's omniscience. For my part, I'm not at all sure that it finally makes sense that God could be rationally certain of all truths. But perhaps it does, and if there were such a being, we've seen no proof that She would have trouble simultaneously satisfying (1) - (3).

Now I don't want to explore the tenability of supposing that a godlike being could rationally be certain of its own rational perfection. For in any case, it seems to me that we cannot simply sidestep our problem by investigating ideal rationality with reference to the beliefs of such a being. After all, a central component of epistemic rationality is having beliefs appropriate to incomplete information. A godlike agent's credences would presumably simply mirror the facts--the agent would be certain of all the truths, have zero confidence in all falsities, and have no intermediate degrees of belief at all. Thus such a model would tell us nothing about a central component of epistemic rationality--the sort of component that's in part captured by something like (2). A useful model of epistemic rationality cannot simply collapse rational belief into truth.

Might there be a non-omniscient ideal agent who could yet be sufficiently free of rational self-doubt to satisfy (1) - (3)? If not, we'd have an argument that rational perfection required factual omniscience. This would, I think, be quite a surprising result. As noted at the outset, rationality seems to be a notion designed in part to abstract from well-informedness. We certainly don't see ordinary cases in which a person lacks information as indicating any sort of lapse in rationality. Yet the suggestion here is that although each of (1) - (3) seems to be aimed at capturing some aspect of thinking well, and not at being well-informed, the three principles together require factual omniscience for their joint satisfaction.

I don't have any argument that this is so. On the surface, what would be required to satisfy (1) - (3) would not be, strictly,

omniscience. We've seen that it would require a great deal of rational certainty about contingent matters concerning the reliability of the agent's own cognitive processes. But this does not obviously imply rational certainty about, e.g., the number of stars in the Milky Way. Nevertheless, it may be difficult to define exactly the scope of propositions over which rational certainty is required, since propositions about the agent's cognitive processes may well be evidentially tied to countless others.

[**Consider first how richly the beliefs of an ordinary agent are interconnected. To the extent that I believe my own cognitive processes to be neurologically realized, propositions about my reliability will relate evidentially to beliefs about the possibility of pharmacological interference with neural functioning, or even the quantum nature of the microprocesses that constitute neural function, etc. And these beliefs will connect to other beliefs about pharmacology and physics, which relate in turn to beliefs about my information sources, be they textbooks or teachers or TV shows, which beliefs relate in turn to countless others. Now for me, this giant web need not be problematic. For I don't have any extreme views about infallibility of my cognitive processes, and thus need not have extreme beliefs about related matters. But for any creature remotely like me, absolute certainty about its own cognitive powers would seem to require absolute certainty about a wide range of contingent matters. Even if one doesn't have worries about the distinction between logical truths and factual ones, there will still be a sense in which our knowledge of logical truths gets ensnared--via reflection on cognitive fallibility--in the web of belief about ordinary factual matters.]

Now perhaps we could find some way of imagining an IRA who could rationally contain the certainty more narrowly. I have no argument showing this to be impossible. But I'm not optimistic. In order not to be required to have extreme beliefs discounting possible cognitive interference stemming from the thousand natural shocks that flesh is heir to, such an IRA would seem to have to be rationally certain, for example, that its cognition was not neurally realized. And it is hard to see how that could be pulled off. Thus at this point, I do not see clearly how to characterize an IRA who's not omniscient, yet who can rationally dismiss all the doubts about its own cognition that would seem

to be required for rational absolute confidence in theorems of logic.

Another way of altering the standard IRA to avoid the tension among (1) - (3) would be to think of an IRA which had no self-doubts because it was completely devoid of beliefs about itself-or, at least, about its own beliefs. After all, it is only when the agent begins to reflect on the possibility of its own epistemic imperfection that the problem seems to arise. Perhaps, instead of imagining an IRA who rationally rejects possibilities of its own error, we could conceive of an IRA who simply never entertains them in the first place.

Again, the question that naturally arises is whether such an agent could be ideally rational. After all, it is not in general rational for an agent to ignore empirical possibilities that bear on the truth of its beliefs. Consider, for example, an ordinary agent who's investigating light, and is absolutely confident that if light isn't a wave, it's a particle, because she's never considered how it could be somewhat like one and somewhat like the other, without being quite like either. In this case, it's not hard to see her absolute confidence in the disjunction (wave or particle) to be the product of a rational failing. And in general, mere failure to consider possibilities does not make it rational to believe as if those possibilities didn't exist.

Perhaps a closer analogy to the case of ideal agents and self-conscious beliefs is that of completely trusting, e.g., visual perception, and ignoring the possibility that things aren't as they appear. Even an agent who had never, e.g., seen a straight stick appear bent in water would not be rational in having absolute confidence that the world was the way it looked. We might think that such an agent was by default entitled to believe that the world was the way it looked, but not that it was entitled to absolute certainty.

Moreover, it's not clear that rationality even applies to an agent who no capacity to reflect on its own beliefs, and has never considered the possibility that it might be mistaken. Some (Davidson maybe) might say that such an agent doesn't even have beliefs. I'm not inclined to go that far, but it does seem strained to call an agent ideally rational if the agent hasn't got the inclination or capacity for any critical reflection on

its beliefs at all. After all, critical reflection on one's beliefs is not just something peripheral to rational belief-management--it seems to be a central component of what it is to believe rationally.

For these reasons, it seems at least doubtful to me that we should consider an agent who doesn't reflect at all on its own beliefs as ideally rational. So I don't think that the tension among (1) - (3) can be avoided by considering unselfconscious but ideally rational agents.

I won't take a stand here on whether (1) - (3) are, in the end, jointly satisfiable, either by an omniscient God or by some lesser being who falls short of complete omniscience. But at this point, I don't see a way of imagining an idealized agent who satisfies (1) - (3) and also can serve as a useful model for studying the question of how non-extreme degrees of belief should be constrained by logical structure. So I'd like to turn now to examine the question of what implications the tension among (1) - (3) has for the study of formal models of rationality.

7. Rational Ideals without IRAs

The suggestion that rationality might require violating coherence raises another question. What should we say about the agent--perhaps an imaginary agent with unlimited cognitive powers but incomplete information--who does take self-doubt into account appropriately, and thus violates coherence? There seem to be two possibilities. First, one could say that, insofar as such an agent achieved the best possible beliefs given her evidence, the agent would be ideally rational. On this view, one would acknowledge that an ideally rational agent might be probabilistically incoherent.

Would this amount to denying that probabilistic coherence was a rational ideal? I think not. We're familiar with other ideals that operate as values to be maximized, yet values whose maximization must in certain cases be balanced against, or otherwise constrained by, other values. In theory choice, simplicity and fit with the data are plausible examples of balancing. In ethics, promoting well-being and respecting rights may illustrate a different sort of constraint. So we can

still see coherence as a rational ideal once we understand its force to be subject to a *ceteris paribus* clause which makes reference to conditions such as respecting evidence and making 1st-order beliefs sensitive to 2nd-order beliefs.

Another option for describing the agent who violates coherence in order to respect other rational desiderata is to deny that its beliefs are ideally rational. There is, I think, something to be said for this option, insofar as it seems that such an agent's beliefs, even if they're the most rational that any agent in the same evidential circumstances could form, are still defective in a clear way: they do not completely respect logic. I at least have more inclination to think this about the incoherent agent than I do about the agent who believes in a theory that's less than maximally simple when the simplest theory doesn't fit the data. Simplicity is a sign of truth, as is data-fit. So the importance of simplicity is derivative, and it's perfectly possible that the goal of truth can be reached without maximizing simplicity. But that's not clearly true of probabilistic coherence. It seems that giving full credence to logical truths is constitutive of our ultimate epistemic goal in a way that picking the simplest theory is not.

On this second view, perfect rationality simply cannot in general be achieved by an agent with limited evidence and unlimited cognitive powers (though perhaps it could be achieved by God). This view is similar to a view of moral perfection that recognizes the existence of real moral dilemmas: circumstances in which even the agent who performs the morally best possible action still acts wrongly. But it seems to me that the position is more plausible in the epistemic case, because it is more plausible to divorce rationality from "ought"-implies-"can" principles. [**Leibniz example?]

Now I'm not sure that the difference between these two views is much more than verbal. One may see ideal rationality as forming the best possible beliefs given one's evidence, or one may see it as perfectly exemplifying all rational ideals. On either way of seeing things, the agent in question forms the best beliefs possible given her evidence, but in doing so must sacrifice perfect attainment of at least one rational ideal. So on either of these views, there turns out to be a strange way in which the mere possibility of epistemic misadventure implies an actual epistemic imperfection.

The interaction between first- and second-order beliefs thus results in something that might be called Murphy's Law for epistemology. The usual version of Murphy's Law states that if it's possible for something to go wrong, it will. The epistemic cousin says that if it's possible that something has gone epistemically wrong (more specifically, if it's possible that I've made a mistake in thinking about T), then something has actually gone epistemically wrong (my belief about T falls short of some rational ideal). For either I'm certain of T, in which case my belief fails to reflect appropriately the possibility that I've made a cognitive error, or I'm uncertain about T, in which case my belief fails to respect logic.

What implications does this have for our theorizing about formal conditions on rational belief? If we agree that all the rational ideals cannot be simultaneously realized by a non-omniscient agent, can we still use idealized agents in thinking about how logic should constrain rational belief? And might this sort of methodology still support taking probabilistic coherence as an epistemic ideal?

I think that we may continue to use idealized agents in studying formal conditions on rational belief. One way to do this is simply to ignore the fact I've been harping on: that the standard idealized agent is violating certain strictures about taking self-doubt into account. We could also, more self-consciously, suppose that an agent was cognitively unlimited, in the sense that it could achieve probabilistic coherence, but then stipulate that either (a) it didn't have any second-order beliefs, or (b) it was certain that it was ideally rational, or (c) it didn't take the possibility of its rational imperfection as a reason to be less than fully confident of logical theorems. Having conceived of our agent in any of these ways, we could then consider arguments that such an agent should have probabilistically coherent beliefs. In one of these ways, it seems to me that we could still run standard arguments based on rational constraints on preferences, or based on invulnerability to Dutch Books.

If we do this, we will have to understand what we are doing in a way that departs from the standard way in which people have thought about the idealized agents they've imagined. We can not, in these cases, think of the imaginary agent as ideally rational. For the agent would be irrationally ignoring or

rejecting epistemically significant possibilities, or failing to take them into account rationally in adjusting its beliefs. Nevertheless, the fact that one is not considering the agent as ideally rational does not, I think, undermine its value as a device to help think about a particular dimension of rationality: how logical structure should constrain degrees of belief.

This can be seen by reflecting on the purpose of imagining idealized agents. The purpose of the idealization is in part to abstract away from contingent human cognitive limitations, and thus to open up the possibility--which is closed off for agents such as us--of attaining coherence. And the idealization should also abstract away from other interfering factors. For example, a Dutch Book argument may assume that the imagined agent values money linearly, and exclusively. The purpose of this assumption is not it's rational to value money this way--the purpose is just to isolate one central way in which beliefs and preferences relate to one another. Now if I'm right, it turns out that one thing that can interfere with an agent's beliefs respecting logic completely is the sort of (rational) self-doubt we've been examining. In stipulating away considerations of (even rational) self-doubt, we create a situation in which the logical constraint on belief can be studied in isolation.

It is important to remember that considerations about the beliefs of ideal thinkers should not anyway be thought of as providing a reductive analysis of the concept of a rational ideal. The idea is not that we take a condition to be a rational ideal in virtue of the fact that the condition would be satisfied by an ideally rational agent. So if it turns out that rational ideals are in tension with one another (at least for agents with limited information) we may reasonably allow one rational ideal to be violated in order to study another under limited-information conditions. So the interest of the idealized-agent-based arguments would not be vitiated by acknowledging that the agents involved were not ideally rational. If coherence can be supported by arguments based on this sort of model agent, that tells in favor of taking it as a rational ideal.

Could the envisioned sort of ideal have the right sort of evaluative implications for humans? Once we admit that our

coherent idealized agent is not, after all, ideally rational, does the whole exercise lose its epistemic significance?

I think that once we see the structure of epistemic ideals in the way I've been urging, we can see that this is not a problem. It's always been clear that the sort of evaluative principle in question--e.g., the more coherent, the better--must be understood as subject to a *ceteris paribus* clause rooted in the limitations of an agent's cognitive system. For example, if improving coherence precluded gathering evidence, or required becoming a paranoid schizophrenic, then *ceteris* wouldn't be *paribus*, and one's beliefs would be less rational if she took the more coherent option. What we've seen now is that the *ceteris paribus* conditions must be understood to encompass another dimension. It's not just that our fleshy limitations might impose epistemic costs on maximizing certain epistemic desiderata. It's also that our status as beings with limited information places some epistemic desiderata at odds with others.

So even for us, it still makes sense to say that the more coherent our beliefs are, the better, *ceteris paribus*. But the *ceteris paribus* conditions make reference to other epistemic ideals. And if the only way of achieving the probabilistically correct attitude toward some claim T would involve embracing irrational beliefs about my own logical invincibility, or violating the principle that my beliefs about logic should cohere with my beliefs about the reliability of the cognitive processes behind my beliefs about logic, then adopting the coherent attitude toward T might well render my beliefs less rational.

So: if all this is right, then the tension among epistemic ideals, at least for agents with limited information, requires us to reconceptualize the sorts of ideal agents often considered in studying formal constraints on degrees of belief, but it doesn't undermine their usefulness. And the fact that the ideal of probabilistic coherence may be constrained by other epistemic ideals, and not just by human limitations, doesn't undermine its status as an epistemic ideal.

However, I do worry that other aspects of formal epistemology might not be left undisturbed by the problem I've been discussing. The classic Bayesian view combines a probabilistic

coherence requirement with a claim about how beliefs are informed by evidence. Conditionalization, and Jeffrey's generalization of it, are the two standard formal accounts of how evidence bears on belief. Both of these accounts presuppose probabilistic coherence.

One might, of course, study these accounts of accommodating evidence by the method I've just recommended for studying logical constraints on an agent's simultaneous beliefs. One might employ probabilistically coherent idealized agents, acknowledging that such agents should not be thought of as ideally rational. And I think that this might well be very useful for studying many cases of evidence bearing on belief. It might even allow us to model cases where some evidential sources undermine others. So an ideal agent who employed conditionalization might allow one to model how strongly I should believe that life in Iraq is improving, given that FOX news said that it is, and given certain information about FOX news's reliability.

But I don't yet see how this would allow us to model the way my belief in T should be affected by evidence that Jocko has drugged my coffee. Stipulating probabilistic coherence gives the wrong result: the probability of T, conditional on any evidence at all, will still be 1. The strategy of abstracting away from the conditions imposed by EVIDENCE or INTEGRATION will not work here, since both of those conditions are centrally important determining how evidence about my being drugged affects the level of credence in T it is rational for me to have. So it seems to me that the tension among our epistemic ideals does pose a problem for traditional formal ways of characterizing how evidence bears on rational belief.

I think that this problem might turn out to be difficult to solve, especially formally. This is because the solution would seem to have to respect all three of the principles; and as we've seen, the principles are in some tension with one another. And the correct way of balancing or constraining one ideal by another is likely to prove difficult to capture formally.

I don't want to argue that this problem can't be solved. One might, for example, try the sort of tactic Dan Garber proposed for handling one version of the old evidence problem. Garber thought the problem stemmed from the assumption of logical

omniscience, and so to relax that assumption, he treated certain logical truths metalinguistically, and allowed ideal agents to be less than certain of their truth. So one might try saying that the credence I should have in T is the probability that "T" is true, given that I seem to have a proof of T and know I have been drugged in a certain way.

This might seem to give the right result in a circumscribed local way. But even this type of model presupposes that the agent is probabilistically coherent over a large range of claims--this is needed for the conditionalization-based mechanism to apply. Garber's particular version assumes that the agent is certain of at least all truth-functional tautologies. But I see no reason to think that proofs of truth-functional tautologies should be exempt from the effects of Jocko's drugs.

Moreover, I suspect that this sort of approach would end up divorcing rational belief too sharply from logic. Even if we restrict our attention to T, and suppose that it's not a truth-functional logical truth, the envisioned mechanism would seem to render irrelevant the quality of the agent's reasoning in proving T. Her proof would enter into determining the rationality of her degree of credence in T only as an apparent proof. The fact that T is a logical truth would have no direct impact on the question of how much rational confidence it merits. Whether a certain inferences were logically correct would have no direct impact on the rationality of beliefs supported by those inferences.

The problem is especially clear if we think about cases of much milder doubt. Suppose Cherry and Kelly are thinking about some matter. Cherry, through flawless reasoning, has concluded that P. Kelly, through logical blunders, has also concluded that P. It seems to me that we need to count Cherry's confidence in P as more rational than Kelly's--even if, when Cherry and Kelly consider their reasons for self-doubt, which are perhaps limited to remembering having made mistakes balancing checkbooks, their reasons are equivalent.

Of course, these worries are only preliminary, and it remains to be seen how difficult a problem we're left with. But if the arguments we've been looking at are correct, whatever account we end up giving of the way beliefs should be informed by evidence

will have to take account of the interaction among epistemic ideals that we've been examining: that what it is rational for an agent to believe in general is constrained by what it is rational for an agent to believe about herself.