Knowledge of Our Own Beliefs

Sherri Roush

Must we have more or less accurate beliefs about our beliefs? Many otherwise diverse thinkers have taken this to be a requirement for rational beings (E.g., Brandom (1994, 2000), Davidson (1984), Moran (2001), Savage (1972), Sellars (1963), Shoemaker (1994, 1996), Williams (2004)). In the tradition that defines rationality by means of the axioms of probability the reason for this view is arguments to the effect that a subject who either failed to be certain that he had a degree of belief he did have, or failed to have a degree of belief he was certain he had, would be vulnerable to sure loss.¹ That is, there is a set of bets that such a subject would accept as fair and that would give him a loss no matter how the events he bet on turned out. This kind of vulnerability, called “incoherence”, is according to this tradition what rationality protects us from. I will argue here that contrary to two entrenched arguments for this view sure loss does not follow from failure to have accurate beliefs about our own beliefs. Thus, although mistaken belief about one’s own belief is a failure, it is a lack of knowledge and not a failure of rationality in the sense expressed by the probability axioms. Put differently, defense of the view that knowledge of one’s own beliefs is necessary for a rational being will need to appeal to more than probabilistic coherence.

Perfect knowledge of our beliefs is one way of achieving coherence when we have higher-and lower-order beliefs, but that by itself gives us no guidance about how to minimize the damage if we are not perfect. My arguments against the sure-loss defenses of perfect knowledge of our own beliefs also identify a bridge principle between higher- and lower-order beliefs, which I call “No Gratuitous Interference” (NGI), the following of which completely protects a subject against sure loss if she does have inaccurate beliefs about her beliefs, and thus preserves rationality in the sense discussed here. The standard bridge principle on this topic, known as “Self-Respect”(SR), protects those with perfect self-knowledge but does not protect the imperfect. Since NGI is a generalization of SR, this new principle is suitable for angels as well as humans, and advisable for those who do not know which they are.

¹ In some discussions of knowledge of our own beliefs (e.g., Davidson 1987) the question is whether given that we have a belief in p, we must know what the content or meaning of p is. Here the question is whether for a given p and assuming we know its content, we must know how confident we are in p.
1. Is ignorance of our belief states even possible?²

It might seem as if there should be no dispute over the claim that we are sometimes wrong, or can be wrong, about what we believe or how strongly we believe it. Like the number of pigs in Allegheny County, that I have a particular degree of belief in a proposition is a contingent fact. It could have been otherwise than it is, and surely could be otherwise for all I know. The mere fact that I possess the belief does not give me special knowledge of it. It does not follow from my possessing a book that I know that I possess that book. If it did then I would either have a much more powerful mind or a much smaller storage unit than I do. I might clearly remember being tempted to buy that book, but not now be sure whether I did or not. I might clearly remember owning that book at one time, but be unable to remember whether I gave it away in the meantime. What is special about the case of belief that makes many people think that if I own a belief I must know that I own it?

The answer may be that there is something special about minds. To extend the book metaphor, it is rather as if, like a person who can’t afford the local real estate, the mind lives in its storage unit with all of the boxes open. If a mind wants to know what beliefs it possesses it has only to turn the light on and rummage around a little. Thus some say that while being wrong about past or future beliefs makes sense, in the same way that being wrong about other peoples’ beliefs makes sense, surely I here now could not be wrong about what I here now believe, at least not if I have a mind. I cannot directly see what books or beliefs were here yesterday, though I might know by some other means, but I can see what books and beliefs are and are not here right now. If it is not immediately obvious then I can just ask myself, and dig around a bit.

You could ask yourself, and your answer could be wrong. Empirical psychology supports the common impression that we know our current mental states better than our past ones, and in many cases know our own mental states better than those of others, and we obviously do know quite enough about our current mental states to get around in the world in daily life, making plans on the basis of predictions of our behavior that depend on our own current beliefs and desires. However, we also have dark corners. Even we who do not regard ourselves as racist or sexist, who would never assent to the claim that minorities or women are by that fact less qualified, can be exposed through experiments to have implicit bias, for example through our overchoosing some job candidates’ resumes over others with no distinction in the profiles except racial association of the name. (Bertrand et al., 2004)

If willingness to act on p is any part of what it is to believe p, then such studies give evidence that we were wrong when we claimed we did not have racist beliefs. It is not that a past self – the one who disavowed racism – was wrong about what a later self – the one undergoing the

² This section is a motivating discussion that is not intended as a survey of even the most important views about self-knowledge of our beliefs.
psychology experiment – believed. There is no reason to suppose the belief changed over that time. It merely took time to collect evidence about what was a stable disposition throughout the story. Thus, one surely can be mistaken about one’s current beliefs, and there can be evidence of it. If the mind were to be master of its own house, its house would have to be as small, inert, and indifferent as a storage unit, and, for good and ill, most minds do not meet these criteria.\(^3\)

If it is conceptually possible to be wrong about one’s own belief states then we should not take sincere assertion of \(p\) as a perfect indicator that one believes \(p\), but we do not. This is easy to see through familiar scenarios that end with the statement “You don’t really believe that.” Imagine a man charged with racketeering, and guilty of it, who has nevertheless successfully avoided his wife ever witnessing anything illegal. Under interrogation the wife quite correctly asserts that she has never seen anything out of order, and quite sincerely asserts that her husband is just a businessman. The astute detective smells weakness, though, stares into her eyes, and says of her summary statement “You don’t really believe that, do you?” She does not want to believe that something is wrong, and does not have evidence that she can specify for believing something is wrong, which together explain her sincere assertion. But she also does not – “deep down” we say – actually believe her husband is just a businessman. Things make her suspect otherwise though she cannot put her finger on why. It may be that all cases of sincere assertion without belief involve self-deception, but that would not support an argument that they are impossible.

Whether it is possible to be wrong about one’s belief states can of course be expected to depend on what one takes belief states to be. If degree of belief in \(q\) is strength of a feeling one has about \(q\), then that would set us up to know what our degrees of belief are, provided feelings are introspectable. Historically, some have thought that looking inward on oneself was not only a source of knowledge but even infallible, but we do not need to assume infallibility in order to understand the logic of this view. The idea is that it is the mind that possesses the feeling, and surely to feel \(q\) is sufficient for feeling that one feels \(q\). Maybe the second-order feeling is not precise, but how can a mind have a feeling it is not (able to become) aware of or sense? Here the intuition comes not so much from a picture of the mind as it did above, but from what feelings are supposed to be. Whatever they are, feelings, hence beliefs, are not like books. The mind does not know one of them by rummaging around and lighting on an object. This view would also explain the widespread impression that one has access to one’s own beliefs of a sort that others do not have, access that has a quality of immediacy. It is easy to see these claims supporting a picture in which having beliefs at all involves having accurate beliefs about those beliefs. How could there be a belief, that is, a feeling, there if the mind that has it could not sense it?

---

\(^3\) The non-technical views cited above that take accurate beliefs about our beliefs as required for rationality can accommodate violations as pathologies. However, that interpretation becomes strained in the case of implicit bias because if what empirical psychology suggests is right then bias in our judgment is pervasive across both subject matters and people. Thus, although morally pathological it would be statistically normal.
Like many probabilists, I follow Frank Ramsey (Ramsey 1926, Armendt 2006) in rejecting the idea that a degree of belief is a feeling, because it is insufficient for acknowledging the causal efficacy that beliefs have on our behavior. Rather, it is part of what belief is that it disposes one to act. Indeed on the Ramsey view having a degree of belief in q is fully identified with being disposed to act as if q is true with a risk of gaining and losing things of value to one that is proportional to that degree; one is disposed to put something at stake on q vs. not-q. To discover what one’s degree of belief is in q by self-inspection would involve not asking oneself how lively q feels, nor asking whether one would sincerely assert q, but asking oneself to what extent one would be willing to act on one’s confidence in q.

But asking oneself questions would not be the only way to investigate what one’s degree of belief is. Actual behavior putting something at risk if q is false is another kind of evidence. No finite set of either kind of evidence will imply that one has a given degree of belief, since a disposition is a regularity of response toward all possible opportunities (that are relevant, or likely, or similar to the actual world, or some such qualifier), but if beliefs are dispositions to act then it is possible in any given case for evidence from actual behavior to be a more reliable indication of one’s degrees of belief than evidence coming from conversations with oneself. This view of belief makes good sense of the impression that the experiments showing implicit bias are uncovering something about our beliefs. Whatever we may say about our beliefs, deciding among job candidates can be more revealing of those mental states because in such decisions we stand to benefit or lose depending on whether we are right or wrong about who is most competent. So we act on the basis of what we really think.

Though Leonard Savage thought of degree of belief roughly the way that Ramsey did, as a state of mind that manifested itself in extravertal behavior (Savage 1972, ), he regarded the answer the subject gives to the hypothetical question how he would bet or act as “just the right one” for “the theory’s more normative interpretation as a set of criteria for us to apply to our own decisions”. (Savage 1972, XX) This is to take it as a norm of rationality that your actual degree of belief (thus betting odds) coincide with what you think it is, and is partly based on the sure-loss arguments discussed below. But the view is not pertinent to the current discussion of whether being wrong about one’s degrees of belief is possible. Savage most definitely thought it was, and that since reports of how one would bet can differ from one’s actual dispositions to bet, the interrogation about hypothetical behavior was a compromise between economy and rigor on the empirical question what a subject’s degree of belief actually is. (Savage 1972, xx)

Savage observed these distinctions but the normative claim that a subject should know what his beliefs are is often conflated with the descriptive claim that he always does via the metaphor of announcement of odds. This accounts for the frequent response on the part of probabilistically well-educated philosophers to the idea that we might not have perfect knowledge of our beliefs, that begins with a blank stare and continues: You’ve just announced your odds! How could you not know what they are? First, having odds does not require announcing them, but second, in the context of odds, announcing is ambiguous between
reporting and doing. It is verbal behavior, but it suggests there is an audience ready to accept those bets. If so, then one who has posted odds is not merely reporting how he thinks he would bet; he is on the hook. But this identifies what we say about our beliefs with what is true of them only by equivocation.

Even if announcement of odds means actually betting, and even though actual betting is better evidence for one’s disposition to bet than merely telling yourself how you would bet, being on the hook on an occasion is not the very same thing as your disposition to bet. You might have made an announcement incompetently relative to your dispositions by making a mistake in the math associated with the stakes in a particular game, or accidentally before you had finished the math. Your actual betting behavior arises out of your dispositions to bet, in conjunction with circumstances and opportunities, but it is not the same thing.

Once we have the idea of a degree of belief as a disposition to act, the possibility of being wrong about our beliefs is easy to comprehend. If it is hard to imagine a belief whose owner does not believe he has it, it is not hard to imagine a disposition to act that a person does not believe he has. More than once in human history a sincere belief that “I would never do that” has been followed sooner or later by the person doing just that. In some cases the disposition will have changed in the meantime between report and act, but it need not; it is enough if the conditions changed in such a way to activate an existing disposition unknown to the subject.

We can imagine others having been in a position to predict the behavior the subject did not expect of himself. As a disposition to act, a belief becomes more analogous to an intention. We seem to find it easy to see when others are not aware of their real intentions, and we have probably all been in situations where it was natural to speak of not knowing whether our own intentions were good.

For Ramsey the picture of a bettor and a bookie looking to score was merely a colorful metaphor. Empirical evidence of one’s dispositions to act can come from any action. The usefulness of the metaphor comes from the fact that every action in which something is at stake can be represented as a gamble. In crossing streets I bet my life that cars cannot move at the speed of light, and, more mundanely, that I’ve looked in the right direction for the country I’m in. These kinds of bets are made without announcements or reports, but we still stand to gain or lose depending on our actions.\footnote{Some resist the betting interpretation of these acts, with the idea that it would be crazy to bet one’s life on an empirical proposition. In fact we do that every day, and we are startled at the idea that because we normally suppress awareness of it, in the way a person who re-locates to an earthquake-prone area and worries about it eventually stops thinking about the risk. Once awareness arises it is common to either suppress it again or else realize that this kind of bet is not taken in isolation. If I never bet my life on matters like cars not traveling the speed of light, I would never cross the street and I would never go anywhere at all. For most people that would not be a life worth having, so the beliefs and utilities balance out.} Thus even though not all dispositions to act are dispositions to lay down money with a bookie, we can imagine them as dispositions to bet, and thereby imagine behavioral evidence about a subject’s degree of belief in \( q \) taking the form of actual betting on \( q \). This simplifies the theoretical discussion, and the quantitative
representation is an efficient way to make qualitative comparisons. Betting at odds of x: 1-x on q would be taken as evidence that the subject has odds of x:1-x, that is, degree of belief x in q.

In the subjective probabilistic view of rationality, rational degrees of belief obey the probability axioms, so the fact that such a subject has degree of belief x in q is expressed by “P(q) = x” where P is the subject’s personal probability function P. These probabilities are possessions of the subject, hence are called “subjective”, but an internalist picture is not mandated by this use of probability. A subject may or may not have introspective access to what a given degree of belief of hers is. A non-extreme degree of belief expresses uncertainty without the subject needing an attitude at all about what her uncertainty is. She may or may not have second-order degrees of belief expressing claims about the first-order degrees of belief, but if she does she need not have introspective access to them either. Notably, she can and standardly does show her appreciation that one claim, q, supports another, p, by revising her degree of belief in p accordingly (e.g., by conditionalization), and without beliefs about that support relation or about her beliefs. In this the framework stands in marked contrast to philosophers who argue for metaknowledge of our beliefs on the basis of a claim that belief revision requires us to have knowledge of our beliefs (e.g., Davidson 1984, Williams 2004: 208, Shoemaker 1994: 281-286, 1996: 33-34).\(^5\)

In fact many of the founding fathers of the probabilistic rationality view were quite opposed to the intrusion of higher-order probabilities or beliefs. These were perceived to be philosophically suspect and mathematical trouble, and it is unclear which was the cart and which was the horse in their arguments.\(^6\) Whether higher-order probabilities are legitimate or need to be reigned in, there is no difficulty representing them since they are a composition of functions. Thus, if a rational subject does have a degree of belief about her own first-order degree of belief, then her second-order degree of belief y in her having first-order degree of belief x in q is expressed by a second-order probability: P(P(q)=x) = y. She has degree of belief y that she has degree of belief x in q.

For our current question we need to imagine behavioral betting evidence of the subject’s degree of belief about her degree of belief in q, but it is easy to see what form this must take. Evidence of her degree of belief about her degree of belief in q would take the form of her bets concerning how she would bet on q. The conceptual possibility of being wrong about one’s own degrees of belief is thus secured by the fact that she could lose the bet about how she would bet on q, by betting differently on q than she bet that she would.

---

\(^5\) One might think that it is no surprise that belief revision by conditionalization can be done without second-order beliefs since conditionalization is not responsivenes to reasons in a full-bodied sense; after all, no deliberative (or even causal) process is portrayed in the probabilistic representation. Prima facie it would not be inconsistent to add to the picture stipulations about what the process should be like, but to do the job of forcing the subject to have higher-order beliefs without begging the question the additional conditions likely would not take the form merely of the subject having degrees of belief in (probabilities for) more propositions.

\(^6\) See Skyrms (1980) for a survey of and response to their objections.
One might think that this possibility is idle since there would be no way to settle the bet she made about how she would bet on q. The only behavioral way to investigate whether she was right about this would be to ask her to bet on q, and this behavior would not be probative because we could not rule out the possibility that this second bet was strategic. She might have bet what she did on q just in order to win her previous bet on what she would bet on q. If she were smart wouldn’t she always do that?

We cannot expect the evidence for our dispositions (degrees of belief) to be infallible whether it comes from behavioral manifestations or introspective conversations. However, we can take steps to address the possibility of strategic betting in the operational procedure described for verification. We can make the reward for getting it right about q much higher than the reward for getting it right about the way she would bet about q, and the reward for the latter small. This reward structure gives her the incentive to bet at the first-order, that is, on q, in accord with the degree of belief she really has in q. (Skyrms 1980) Since she got no new evidence about q between the two bets, we can assume that the degree of belief she manifests in the second bet is also the one she had when she bet about what her degree of belief in q was.

How can the Ramseyan picture explain the impression that our knowledge of our own beliefs is better and more immediate than our knowledge of others’ beliefs or their knowledge of ours? Introspection is immediate in some sense and can only be done on oneself, so the existence of such an ability would explain the intuitions of asymmetry between the types of access the 1st and 3rd person have to information about belief states. This asymmetry of access could also provide some explanation of why we typically know more about our own beliefs than others do.

The Ramseyan view does not require introspective access but the existence of introspective evidence for belief is not incompatible with the view that belief is a disposition to act. What would be incompatible is the view that introspection and sincere assertion are the only kinds of evidence. The Ramseyan view of belief gives the means to tell a fuller story. It is possible that at least some of our knowledge of our own mental states, including our dispositions to act, is gained in the same manner as our knowledge of others’ mental states – via behavioral data, our observations of our own actions. This view is supported by a good deal of current cognitive science (Carruthers 2011), and the view that this is our primary way of knowing our own minds goes back at least as far as Gilbert Ryle (Ryle 1949: 155-6).

Whenever we do use behavioral evidence to know our own minds, an asymmetry in our knowledge of beliefs is introduced by the fact that in our own case we have a lot more empirical evidence. With the exception of conjoined twins, a human being spends more time in her own company than she does with any other individual. She thus has anywhere from more to vastly more behavioral evidence about herself than about any other individual, and than any other individual has about her. The felt immediacy of our knowledge of our mental states, when that feeling exists, might come from the fact that we do not need to be consciously thinking about our behavior in order to be registering or processing information about it, and
that at any given time most of our behavioral evidence about ourselves will already have been processed and our conclusions ready to hand. Thus the Ramsey view of beliefs allows us to add a type of evidence that we might have about our beliefs, and that can contribute to explaining 1st- and 3rd-person asymmetry, without requiring us to deny that there is introspective evidence.

Just as the possibility of self-deception does not imply that we are pervasively self-deceived, the possibility of inaccurate beliefs about one’s beliefs does not imply that we are frequently, grossly, or pervasively wrong about them. We are evidently not, as noted above, since we effectively anticipate and plan many of our actions. The upshot of the conceptual possibility of error about some of our beliefs is support for the view that our accuracy about our beliefs in ordinary contexts is a contingent fact, not a necessary consequence of having beliefs at all.

2. Is Ignorance of one’s own belief states compatible with rationality? The Direct Argument

To admit that blindness about some of our beliefs is conceptually and psychologically possible, is not to concede that such ignorance is compatible with being a rational subject, and the latter is usually the point at issue when the possibility of self-blindness about one’s beliefs is denied. I will focus here on a way of asserting this claim in probabilistic terms, and two ways of defending the claim that get little discussion in print because they are widely taken to be a settled matter, obvious to anyone sufficiently trained to follow a two-step and a four-step argument about betting. I will argue that both of these arguments are invalid, and propose an alternative, systematic, way of handling uncertain and not fully accurate beliefs about our beliefs that avoids sure loss and has recognizable and compelling intuitive interpretations.

The natural way for a probabilist to define knowledge of one’s own belief states is through the following conditions:

For every x and every q,

if \( P(q) = x \) then \( P(P(q) = x) = 1 \) \quad \text{Confidence}

if \( P(P(q) = x) = 1 \) then \( P(q) = x \), \quad \text{Accuracy}

Confidence says that if one’s degree of belief in q is x, then one is certain that one’s degree of belief in q is x. Accuracy says that if one is certain that one’s degree of belief in q is x then indeed it is x. 7 I will call these two conditions together “Self-Transparency” (ST). 8, 9

---

7 I follow David Christensen’s (2007) names for these conditions.
8 In these stipulations knowledge of one’s beliefs is defined as infallible and perfectly precise, but in rejecting ST, infallibility and over-precision are not my targets. Using the new principle, NGI, a subject can protect herself from sure loss no matter the level of inaccuracy she has about her beliefs. The point here is not that she be allowed imprecision, but that failure to know what her beliefs are is not a failure of rationality.
Soshichi Uchii (1973) argued that addition of these conditions to the probability axioms yields the natural extension of the probabilistic conception of rationality to the second-order, that is, to degrees of belief about degrees of belief, or probabilities of probabilities, via a sure-loss argument that is repeated regularly and reflexively today.

To show vulnerability to sure loss we must show that there is a set of bets that the subject regards as fair and that would give her a loss no matter how the questions she bet on turned out. In the sure-loss argument for ST we have a subject whose probability function is $P$, and who has degree of belief $x$ in $q$. The statement that she has this degree of belief, $P(q) = x$, is the proposition we imagine her betting on, call it $B$. That is, we imagine her having a degree of belief, $z$, about whether $x$ is her degree of belief in $q$, and $z$ is her probability for the proposition $P(q) = x$; we write this statement $P(P(q) = x) = z$, and now $P(B) = z$. Her gains when $B$ is true and false are as usual, with $S$ the stake:

<table>
<thead>
<tr>
<th></th>
<th>B true</th>
<th>B false</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S - zS$</td>
<td>$- zS$</td>
<td></td>
</tr>
</tbody>
</table>

One way for the subject to violate self-transparency is for $P(q) = x$ to be true while she is not certain of it. That is, $z < 1$ and $B$ is true. Now suppose the stake is $-1$. Her net gain if $B$ is true is $S - zS$, which is $-1 + z < 0$. Her net gain if $B$ is false is $z > 0$. But $B$ is not going to be false because it is a statement of this subject’s degree of belief in $q$ and that is $x$ by assumption. Thus, the subject whose probability function is $P$ does not have a chance at the gain in the column where $B$ is false. If the stake on the bet on $P(q) = x$ is negative, then the subject for whom $P(q) = r$ but who is uncertain that $P(q) = r$ is sure to suffer a loss of $x - 1$.

The other way for the subject to violate ST is for her to be certain that $P(q) = x$ when it is not. I.e., $z = 1$ and $B$ is false. In this case, take a positive stake, $S = 1$. If $B$ is true, then she has a gain of $1 - z = 0$, and if $B$ is false then she has a loss of $z$. But this subject’s degrees of belief about $q$ insure that $B$ is false, so she can only lose. This subject is sure to suffer a loss of $z$.

This argument shows that for a subject with probability function $P$ and who violates self-transparency there is a betting book that she would accept that would give her a loss in all possible worlds in which her probability function is $P$. It is a result that should come as no surprise, since restricting the set of worlds of evaluation to those in which she has this probability function is effectively treating the fact that she has these degrees of belief as a

---

9 One could define self-knowledge using two different probability functions, instead of one function applied to itself as in ST. One of the two functions would play the higher-order role, the other representing degrees of belief at the first order. However, that might be similar enough to one subject having degrees of belief about another subject’s degrees of belief that inaccuracies would escape incoherence in a similar way; there does not seem to be any intuitive reason to think that being mistaken about someone else’s beliefs makes us irrational. I use one probability function applied to itself because this is prima facie the most likely to produce incoherence when there is any misalignment between the two orders.
necessary truth. A subject will be vulnerable to sure loss if she bets even a penny against a necessary truth.

As I have stressed, that a subject has a particular degree of belief is a contingent truth. Two plus two could not have been 5, but our subject’s degree of belief in q could have been different than it actually is, and, as we implicitly grant by imagining this betting scenario, it could actually be different from x for all she knows. The fact that what her degree of belief is in q is settled at the stage of our betting scenario does not change this. Even if a coin is already tossed, it is still sensible to bet on two possible outcomes, as long as the result of the toss remains concealed.\(^{10}\) The fact that we the theorists may legitimately assume we know what the subject’s probability function is does not imply that the subject knows. One way to justify the claim that the subject knows what her probability function is would be to assume that she must know in order to be a rational subject at all, but of course that would be begging the question at issue.

Substituting a concealed coin toss for her degree of belief in the sure-loss argument above brings out the odd quality of that argument. In the substitution the analog to her having degree of belief x in q is the coin coming up, say, heads, H, and the analogs of violating self-transparency are being uncertain that the coin came up heads when it did, and being certain that it came up heads when it did not, i.e., the two ways of being wrong about whether it came up heads. If we assume that the coin indeed came up heads, H, and that this has been concealed from the subject, and as above that her probability function is P, the sure-loss argument above becomes:

One way for the subject to be wrong about whether it came up heads is for H to be true while she is not certain of it. That is, \(z < 1\) and H is true. Suppose the stake is \(-1\). Her net gain if H is true is \(S - z \cdot S\), which is \(-1 + z < 0\). Her net gain if H is false is \(z > 0\). But H is not going to be false because it is a statement of the outcome of the coin toss, and that is H by assumption. Thus, the subject whose probability function is P does not have a chance at the gain in the column where H is false. If the stake on the bet on H is negative, then the subject for whom H is true but who is uncertain that H is sure to suffer a loss \(z - 1\). ...

Necessarily, if an outcome is settled, then a subject who bet anything against that outcome will lose something, but vulnerability to sure loss requires a set of stakes that will make you lose in all possible outcomes. If you have a false belief about the outcome of the coin toss, a bookie who knows the outcome of the toss could exploit your ignorance for his gain, but this loss is due to a lack of knowledge, not a failure of rationality, on your part.

To establish a sure-loss vulnerability in the foregoing way in a subject who is not Self-Transparent, it has to be not just true that \(P(q) = x\), but true in all possible worlds relevant to

\(^{10}\) Thanks to Joe Ramsey for suggesting this comparison.
the evaluation. \( P(q) = x \) is a contingent matter, so, like \( H \), its actually being true does not imply there are no possible worlds in which it is false. Why did we only count as relevant those worlds in which \( P(q) = x \) is true? It is standard procedure when evaluating the subject’s fate in betting that one only evaluates worlds in which she has the degrees of belief, i.e., probability function, that she actually has. Here the subject actually has degree of belief \( x \) in \( q \), so we take it she has this in every world relevant to the evaluation. The reason for this procedure is that her probability function determines the odds she is willing to accept, and we are trying to determine what fate those dispositions to bet will bring her, not what would happen if she had some other dispositions.

However, this rationale for the standard procedure of only considering possible worlds compatible with the subject’s actual probability function supports a different procedure in the second-order case. The set of relevant possible worlds should indeed be ones where she has the odds she uses in the actual world for the questions she is betting on.\(^{11}\) When she bets on \( q \), the odds she is willing to accept are determined by the value of \( P(q) \), but when she bets on whether or not \( P(q) = x \), her odds are determined by a different part of the function \( P \), namely, by the value of \( P(P(q) = x) \). The same rationale that tells us that when she bets on \( q \) we should consider only worlds in which \( P(q) = x \) implies that when she bets on \( P(q) = x \) we should consider worlds in which \( P(P(q) = x) \) equals its actual value, but it gives no reason to restrict the outcome-worlds for the second-order bet to those worlds in which \( P(q) = x \).

Indeed, if we do hold \( P(q) = x \) fixed, that is, true in all worlds relevant to the evaluation, when we evaluate how the subject fares in betting on \( P(q) = x \), we are not treating \( P(q) = x \) as an outcome or random variable, a proposition for which more than one value is possible, in this case the values true and false. If so, then we are not treating it as something that could be bet on, so the Uchii argument is not a betting argument, not a sure-loss argument for Self-Transparency, at all.

In the subjective interpretation of probability that \( P(q) = x \) is a random variable is equivalent to its being possible that the subject does not know whether her degree of belief in \( q \) is \( x \). A random variable can be thought of as a proposition whose probability value is subject to variations due to randomness. What this randomness consists in depends on the interpretation of probability. For example, if one has a propensity interpretation, then the randomness is the chance involved in the set-up for an experiment, e.g. with dice. On the subjective interpretation of probability, it is the "subjective" randomness that results from incomplete knowledge of a quantity; it is epistemic uncertainty. There is no question that in the current context we are using a subjective interpretation, and therefore allowing or denying that \( P(H) = x \) is a random variable is equivalent to allowing or denying it as possible that the subject does

\(^{11}\) This distinction is discussed in slightly different terms by Rachael Briggs (2009) in the course of discussion of a sure-loss argument closely related to the one I discuss in the next section.
not know the outcome, here does not know what her belief is. In not treating $P(H) = x$ as a random variable, the Direct Argument begs the question.

Denying that $P(q) = x$ could even be an outcome or random variable is an option, but one would need an argument for it that addresses the fact that whether one has a particular degree of belief or not is a contingent matter. Sure loss arguments can be made against a subject who fails to be certain of logical truths or who is certain of logical falsehoods, but what justifies this is that the content of the proposition is a necessary truth.\(^{12}\) We can treat those worlds in which the necessary truth is true as the only worlds relevant to the evaluation – hence not treat the necessary truth as a random variable – because there just are no worlds in which these truths are false. Since the content of the claim that a subject has a degree of belief is a contingent matter, the burden is on one making a sure loss argument for Self-Transparency to explain why we should treat it as necessary. It is a necessary truth relative to the subject’s probability function, but resting our argument on that buys nothing. Her probability function as a whole also could have been otherwise.

Holding $P(H)$ fixed when evaluating the subject’s bet on $H$, and letting it vary when she is betting on the proposition $P(H) = x$ will be enforced simply by treating the bets on $H$ and the bets on $P(H) = x$ differently. Dividing up the probability function into parts that she bets with – to be taken as having in all relevant worlds the same values as they have in the actual world – and parts that she bets on – statements about which are true or false depending on the world – may seem suspect since in evaluating coherence we want to know whether a person’s degrees of belief fit together properly. Coherence of a whole function cannot be evaluated a proposition at a time.

It is true that we can only succeed in evaluating the coherence of a set of betting odds if we consider them together. It is also true that evaluating the coherence of a proper subset of a particular subject’s degrees of belief will not be enough to conclude that the subject is coherent. However, doing this second-order sure-loss argument as I will propose need not run afoul of these points. The current argument can be imagined without dividing the probability function into more than two parts, and the division can be made in a principled way, indeed syntactically, as a difference between first- and second-order. Probabilities of the form $P(q)$, where $q$ does not contain any “$P$” will be the first set, and those where $q$ does contain “$P$” are the second. Moreover, we are entitled to assume that the first-order degrees of belief of the subject are coherent, since our question is whether failure of transparency at the second-order about one’s beliefs at the first order introduces incoherence. In fact we must assume the subject is otherwise coherent in order to isolate this question. Thus, if we can see that

\(^{12}\) Even this is dubious since the randomness of a random variable in the subjective interpretation is supposed to be epistemic uncertainty, and propositions whose content is logically necessary can still be epistemically uncertain. The probabilistic conception of rationality is prima facie limited to requiring logical omniscience, a problem beyond the scope of this paper.
inaccurate second-order degrees of belief do not introduce incoherence into an otherwise coherent subject, then we will have certified the coherence of such a subject’s entire set of probabilities, not just a subset of them.

With this understanding, we are imagining a subject betting with the second-order part of her function on the first-order part of her function, with the assumption that the first-order part of her function is coherent, but no further assumptions about the particular values its arguments take. She fails Self-Transparency if she lacks confidence or accuracy about even one of her degrees of belief, that is, about propositions of the form \( P(q) = x \). We ask whether a failure of this sort yields a sure loss. The sure loss argument above fails, for the same reason that taking its analog in the concealed coin toss situation as a sure loss would be a misinterpretation. If, in the first way of failing self-transparency, \( P(q) = x \) is true but \( P(P(q) = x) \neq 1 \), then a negative stake is the only way to argue for a sure loss, but it does not yield a sure loss. In one of the possible outcomes, namely where \( P(q) \neq x \), the subject gains \( z > 0 \). If her bet against \( P(q) = x \) corresponded to a degree of belief in \( P(q) \neq x \) that was greater than 0, then she wins something. That non-actual world in which \( P(q) \neq x \) and she wins something for betting even a penny in its favor is among the possible worlds relevant to the evaluation. A similar point holds for the other kind of failure of transparency.

The view I am advocating has a number of virtues. Intuitively, the difference between the Direct Argument and the correction I am proposing is that the former is posed from the theorist’s point of view, whereas I set up the betting from the point of view of the subject. The latter, but not the former, is in keeping with the subjective interpretation of probability, as noted previously concerning random variables, and with the whole point of the conception of rationality based on subjective probability, that the subject not be judged on the basis of whether her beliefs are true or probable. This is why it makes sense for her to be judged not on whether she loses in the actual world, but whether she could lose in all possible worlds. When we judge her by coherence we find that the subject will actually lose, other things equal, if she does not know what her degrees of belief are, but she will not necessarily lose. This distinction may seem idle, but we will see below that it is not if her questions about what to believe about her beliefs are embedded in a larger context of questions she faces.

3. Self-Transparency and “Self-Respect”: The Indirect Argument

Though the direct sure-loss argument for Self-Transparency fails, there is a bridge principle that immediately implies the Accuracy direction of ST, and for which a sure-loss argument has also been proposed. This is
called simply “Self-Respect” by Christensen (2007), and a two-function version of which was dubbed “Miller’s Principle” by Brian Skyrms (1980). I will sometimes refer to this as the “10th-Character Principle” since it requires on the right hand side a re-inscription of whatever is in the 10th-character position, which happens to be on the left-hand side. RSR says that your degree of belief in H given that your degree of belief in H is x, should be x. We could justify it by saying that the mere consideration that your degree of belief in H is x does not provide any reason to think that your degree of belief in H should be something other than x. RSR implies the Accuracy condition above because

\[ P(P(q) = x) = x \quad \text{Restricted Self-Respect (RSR)} \]

\[ \text{Christensen (2007) has shown that Confidence and Accuracy together imply what I am calling RSR. Notably, RSR alone does not imply Confidence.} \]

\[ \text{RSR is the synchronic instance of Bas van Fraassen’s (1984) Reflection Principle, so Reflection presupposes and implies perfect knowledge of one’s current beliefs.} \]

RSR or variations of it have played a key role in arguments illustrating the usefulness of second-order probabilities, and the logically stronger ST has played that role in arguments urging their triviality. For example, a version of RSR with different functions at the two orders allowed Skyrms to show that the content of what we learn in a Jeffrey conditionalization can be understood as a second-order disjunction of statements of the degrees of belief that changed in the learning, because second-order strict conditionalization on such a disjunction is equivalent to first-order Jeffrey conditionalization. Haim Gaifman used a two-function unrestricted version of RSR as an axiom to construct a theory of higher-order probability.

(Gaifman 1986)

---

13 I call it “Restricted” to distinguish it from a stronger principle, \( P(q/P(q)=x \cdot S) = x \) for S any statement of probability, that I call “Unrestricted Self-Respect” (USR). Even if the sure loss argument for RSR below succeeded, it does not succeed for USR since the conjunction in USR’s condition yields additional possible worlds. Restricted Self-Respect is also distinct from Christensen’s (2007) Moderate Self-Respect, which requires that the more the subject believes the condition, the closer the value on the right-hand side is to x. The principle I will substitute below for RSR allows us to avoid sure loss at any level of departure from RSR’s right-hand side.

14 Assume RSR. Without loss of generality suppose that the only possible values for \( P(H) \) are x and y, and suppose \( P(H) = x \). By total probability, \( P(H) = P(H/P(H)=x)P(P(H)=x) + P(H/P(H)=y)P(P(H)=y) \). Under our assumptions, it follows that \( x = xP(P(H)=x) + yP(P(H)=y) \); to preserve coherence it is sufficient that \( P(P(H)=y) | P(P(H)=x) = x | y \). This means that the Indirect Argument for RSR cannot serve as a full argument for ST.
On the other side, Self-Transparency makes all of the second-order absolute probabilities equal zero or one, thus making them irrelevant to every other proposition, and so, it is assumed, idle, unable to effect any changes in the probabilities of other propositions.\textsuperscript{15} ST is thus a perfect shield from the perceived trouble of second-order probabilities since it acknowledges the existence of the composition of functions that is implied by the probability representation, while (apparently) rendering any hierarchies thus expressed ineffectual and innocuous. Thus, whether we should take it as a requirement that the value of \( P(q/P(q) = x) \) be \( x \), and if so why, is a matter of some interest independently of the question whether we must have knowledge of our beliefs in order to be rational.

Skyrms seems to regard his version of Self-Respect as a useful idealization, having even provided counterexamples to it (Skyrms 1980, 125).\textsuperscript{16} While I think there can be no quarrel that various versions of this principle are valuable for simplifying a representation and isolating questions of interest, many have regarded RSR as a requirement of rationality, on the basis of an argument that violating it makes a subject vulnerable to sure loss. If that argument succeeds then coherence implies RSR, which implies Accuracy, so coherence requires Accuracy at least, but I will argue that the sure-loss argument fails.

The substance of this sure loss argument has been discussed in illuminating ways and in different forms by Christensen 2007, and Briggs 2009, both of them pointing out ways in which the sure loss that is secured by it is weaker than the usual conclusion. I will go further to argue that there is an alternative principle to RSR that avoids sure loss vulnerabilities entirely, even for the subject who fails Self-Transparency, even if the failure is extreme.

Here I present the sure-loss argument in the style that followers of Savage will recognize. Taking

\begin{itemize}
  \item H: The coin lands heads.
  \item G: \( P(H) = x \)
\end{itemize}

we suppose that

\[ P(H/P(H)=x) = y \text{ for some } y > x, \]

that is, that the subject violates RSR, and

\[ P(P(H)=x) = z > 0, \]

\textsuperscript{15} The Savage-Woodbury “collapse” argument (Savage 1972, XXX) that since any statement of second-order probability can be reduced to a first-order statement it is thereby trivial or epiphenomenal, does not, or need not, assume RSR. It depends only on the fact that total probability relates the first- and second-order thus:

\[ P(H) = P(H/P(H)=x)P(P(H)=x)) + P(H/P(H)\neq x)P(P(H)\neq x). \]

That argument does not imply ST and is not intended to.

\textsuperscript{16} His counterexample employs an assumption that could be written as a conjunct in the condition, and so as an instance of USR rather than RSR. See fn. 13.
that is, that the subject allows it as at least possible that the condition \( P(H) = x \) is fulfilled.\(^ {17} \)

To represent a conditional probability in betting terms we use a called-off bet. For the conditional probability \( P(H/G) \), the bet concerning \( H \) will only be in force if \( G \) holds. If \( G \) does not hold then the bets involving \( H \) are off. In our case \( G \) is \( P(H) = x \). That is, \( G \) is the claim that the subject’s degree of belief in \( H \) is \( x \). The subject has odds on both \( H \) and \( P(H) = x \), but the bet the conditional probability tells us he makes on \( H \) is called off if the other matter he has odds on, \( P(H) = x \), turns out false.

To see how the commitments expressed in our assumptions might turn out for the subject, we write a table with columns for all of the possible outcomes and with stakes, \( b_n \). \( H(\omega) \) is 1 if \( H \) is true, and 0 if \( H \) is false, and similarly for \( G(\omega) \), which is 1 if \( P(H) = x \) and 0 if \( P(H) \neq x \).

<table>
<thead>
<tr>
<th>Outcomes ---&gt;</th>
<th>H and G</th>
<th>-H and G</th>
<th>-G</th>
</tr>
</thead>
<tbody>
<tr>
<td>( G(\omega)b_2(H(\omega) - y) )</td>
<td>( b_1(1-y) )</td>
<td>-( b_1y )</td>
<td>0</td>
</tr>
<tr>
<td>( b_2(G(\omega) - z) )</td>
<td>( b_2(1-z) )</td>
<td>( b_2(1-z) )</td>
<td>-( b_2z )</td>
</tr>
<tr>
<td>Total each outcome</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

There are only three columns because in case \( G \) is false, the bet corresponding to the conditional probability is called off, so it does not matter whether \( H \) is true or false. With the proposition \( P(H) = x \) represented as “\( G \)”, this looks like an ordinary situation. But \( G \) is a statement of probability, hence a statement of the subject’s betting odds on \( H \). This statement will be true in the first two sets of possible worlds, and in those worlds the odds \( G \) indicates will contribute to the subject’s wins and losses. In those worlds, in addition to being willing to bet at odds \( y:1-y \) on \( H \), she is also willing to bet at odds \( x:1-x \) on the same proposition. That commitment must be listed in the table along with the other odds, yielding:

<table>
<thead>
<tr>
<th>Outcomes ---&gt;</th>
<th>H and ( P(H) = x )</th>
<th>-H and ( P(H) = x )</th>
<th>( P(H) \neq x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( b_3(H(\omega) - x) )</td>
<td>( b_3(1-x) )</td>
<td>-( b_3x )</td>
<td>------</td>
</tr>
<tr>
<td>( G(\omega)b_2(H(\omega) - y) )</td>
<td>( b_1(1-y) )</td>
<td>-( b_1y )</td>
<td>0</td>
</tr>
<tr>
<td>( b_2(G(\omega) - z) )</td>
<td>( b_2(1-z) )</td>
<td>( b_2(1-z) )</td>
<td>-( b_2z )</td>
</tr>
<tr>
<td>Total each outcome</td>
<td>( b_3(1-x)+ b_1(1-y)+ b_2(1-z) )</td>
<td>-( b_3x+ -b_1y+ b_2(1-z) )</td>
<td>-( b_2z )</td>
</tr>
</tbody>
</table>

\(^ {17} \) I represent the subject’s beliefs as probabilities rather than merely as credences, even though whether those credences are coherent is the point at issue, because I am assuming they are coherent until some specific hypothesis about the subject makes them otherwise.
Accepting two different sets of odds on the same proposition can come to no good. Following Teddy Seidenfeld, who endorses this argument fully and thinks that Savage had it in mind as obvious, we can see the subject described by this table as vulnerable to loss in all possible outcome-worlds, if we assign values 1, (y-x)/2, and -1 to the stakes b₁, b₂, and b₃ respectively. This gives a negative payoff in every column, as calculated in the final row below.

<table>
<thead>
<tr>
<th>Outcomes ---&gt;</th>
<th>H and P(H) = x</th>
<th>-H and P(H) = x</th>
<th>P(H) ≠ x</th>
</tr>
</thead>
<tbody>
<tr>
<td>b₃(H(w) − x)</td>
<td>b₃(1-x)</td>
<td>-b₃x</td>
<td>-------</td>
</tr>
<tr>
<td>G(w)b₁(H(w) − y)</td>
<td>b₁(1-y)</td>
<td>-b₁y</td>
<td>0</td>
</tr>
<tr>
<td>b₂(G(w) − z)</td>
<td>b₂(1-z)</td>
<td>b₂(1-z)</td>
<td>-b₂z</td>
</tr>
<tr>
<td>Total each outcome</td>
<td>x-y + (y-x)(1-z)/2</td>
<td>x-y+ (y-x)(1-z)/2</td>
<td>(x-y)z/2</td>
</tr>
</tbody>
</table>

x-y is negative, 1-z is no more than 1, and (y-x)/2 < y-x, so the first two outcomes both have losses under this assignment. x-y<0 and z>0, so the third outcome is also a loss. Thus there are stakes at which in every relevant possible world the subject loses. An analogous argument can be made for y < x, z > 0.

How does an argument style that did not work to defend ST, work to defend something that implies it? We saw that what undermined the sure loss arguments concerning the unconditional probabilities, P(H) = x and P(P(H)=x) = 1 was that the non-actual worlds in which P(H) did not equal x – and the subject who was uncertain that it was x might win – exist and had to be counted among the relevant possible worlds if the subject’s having a particular degree of belief was to be treated as a random variable. The conditional bet by its structure (apparently) removes those worlds from the set of relevant possibilities. The possible worlds in which the subject who has odds on H that are not equal to x could win or break even, also happen to be worlds in which the conditional bet is called off.

Neat as this is, the previous table is misleading in listing the bet on H at odds x:1-x on a par with the other bets, for while the odds implied by our assumptions:

\[ P(H/P(H)=x) = y \text{ for some } y > x, \]
\[ P(P(H)=x) = z > 0, \]

describe actual dispositions of the subject, neither of these assumptions nor their conjunction determines an actual value for P(H). Nothing in our assumptions allows us to say more about the odds x:1-x than that it is possible that the subject has them. Thus, their difference in status from the odds listed in the left column needs to be flagged more prominently.

Yet such a table will still not be complete, for though the subject does not have a commitment to odds of x on H in worlds of types three and four, she does have some commitment or other
about H in each of those worlds; we will call those odds u:1-u. Finally, the subject also has some value or other for P(H/P(H)≠x) in every possible world, and that must be registered in the table to see the full picture. It is a disposition with regard to the complementary conditional bet to the one in our original assumptions, the probability of H on the catch-all, that is, on the assumption that P(H)≠x. We can see via total probability that lack of a value for this term is why our two assumptions do not determine an actual value for P(H):

\[ P(H) = P(H/P(H)=x)P(P(H)=x) + P(H/P(H)≠ x)P(P(H)≠ x) \]

If we assume RSR, then P(H/P(H)=x) = x for all x, and that determines the value of P(H/P(H)=u) for all u not equal to a specified x; the value is u. However, we were supposed to be arguing for RSR, not assuming it. Our assumption that P(H/P(H)=x) = y > x does not determine the value of P(H/P(H)≠x) – other things equal, likelihoods are independent of each other, and the assumption puts no condition on the relation of y and x except that y be greater – so we must record this likelihood explicitly in the table in order to allow that as far as we know it can take any value greater than x. The odds corresponding to this conditional probability we represent here as v:1-v.

With all of the needed terms spelled out properly, the table looks like this:

<table>
<thead>
<tr>
<th>Outcomes ---+--</th>
<th>H and P(H) = x</th>
<th>-H and P(H) = x</th>
<th>H and P(H)= u ≠ x</th>
<th>-H and P(H)=u ≠ x</th>
</tr>
</thead>
<tbody>
<tr>
<td>“b_n(H(ω)−x)”</td>
<td>b_3(1-x)</td>
<td>-b_3x</td>
<td>b_5(1-u), etc.</td>
<td>-b_5u, etc.</td>
</tr>
<tr>
<td>Actual odds  ↓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1-G(ω)b_4(H(ω)−r))</td>
<td>0</td>
<td>0</td>
<td>b_4(1-v)</td>
<td>-b_4v</td>
</tr>
<tr>
<td>G(ω)b_1(H(ω)−y)</td>
<td>b_1(1-y)</td>
<td>-b_1y</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>b_2(G(ω)−z)</td>
<td>b_2(1-z)</td>
<td>b_2(1-z)</td>
<td>- b_2z</td>
<td>- b_2z</td>
</tr>
<tr>
<td>Total each outcome</td>
<td>b_3(1-x)+ b_1(1-y)+ b_2(1-z)</td>
<td>-b_3x + -b_1y + b_2(1-z)</td>
<td>b_5(1-u) + b_4(1-v) − b_2z</td>
<td>-b_5u -b_4v − b_2z</td>
</tr>
</tbody>
</table>

Recognizing the catch-all conditional probability does not by itself make the sure-loss vulnerability go away. We can make a subject with the commitments expressed in this table lose in the first two sets of worlds and on the bet with payoff –b_2z as before, and as long as the coefficients b_2 and b_3 on the additional terms in the last two sets of worlds are chosen both with absolute values much less than b_2, i.e., much less than (y-x)/2, then wins on those bets cannot compensate for the loss represented by –b_2z.

The more explicit table also supports the view that the subject can avoid sure loss by taking the conditional bet with the confidence indicated by whatever stands in the 10th character position, that is, by following RSR:
I believe that H.

Outcomes

That a subject following NGI is not vulnerable to sure loss can be seen in the following table:

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>H and P(H) = x</th>
<th>-H and P(H) = x</th>
<th>H and P(H)= u ≠ x</th>
<th>-H and P(H)=u ≠ x</th>
</tr>
</thead>
<tbody>
<tr>
<td>“bₙ(H(ω) – x)”</td>
<td>bₙ(1-x)</td>
<td>-bₙx</td>
<td>bₙ(1-u), etc.</td>
<td>-bₙu, etc.</td>
</tr>
<tr>
<td>Actual odds ↓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1-G(ω)b₄(H(ω) - q))</td>
<td>0</td>
<td>0</td>
<td>b₄(1-u)</td>
<td>-b₄u</td>
</tr>
<tr>
<td>G(ω)b₁(H(ω) – x)</td>
<td>b₁(1-x)</td>
<td>-b₁x</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>b₂(G(ω) – z)</td>
<td>b₂(1-z)</td>
<td>b₂(1-z)</td>
<td>- b₂z</td>
<td>- b₂z</td>
</tr>
<tr>
<td>Total each outcome</td>
<td>(b₃+ b₁)(1-x)+ b₂(1-z)</td>
<td>-(b₃x +b₁)x+ b₂(1-z)</td>
<td>(b₄+ b₅)(1–u) – b₂z</td>
<td>-(b₄+ b₅)u – b₂z</td>
</tr>
</tbody>
</table>

In none of the worlds represented by the columns does the subject have more than one set of odds on H, so no coefficients will lead to loss in all possible worlds.

However this does not imply that following RSR is the only way to avoid incoherence. With the full table we can see that there is another safe way for a subject to specify a value for P(H/P(H) = x). Imagine a subject who has a policy of regarding the statement that she has a particular degree of belief in H as irrelevant to whether H is true. Thus, whichever world she might be in, and for all x, her disposition toward P(H/P(H)=x) is just the same as her disposition toward H, whatever that disposition toward H might be in that world. She would be following a strategy of regarding her mere belief about what her belief in H is as irrelevant to what her belief in H should be. This is expressed by the following alternative principle to RSR:

P(H/P(H)=x) = P(H) No Gratuitous Interference (NGI)

We justified RSR above by saying that merely being confident that I have a given degree of belief in H is not a reason to have a different degree of belief in H. NGI follows the thought that merely being confident that I have a given degree of belief in H is also not by itself a reason to have that degree of belief in H. If I am a probabilistically rational subject I have reasons for whatever degree of belief I do have in H, but that I have that degree of belief in H is not by itself a reason.

That a subject following NGI is not vulnerable to sure loss can be seen in the following table:

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>H and P(H) = x</th>
<th>-H and P(H) = x</th>
<th>H and P(H)= u ≠ x</th>
<th>-H and P(H)=u ≠ x</th>
</tr>
</thead>
<tbody>
<tr>
<td>“bₙ(H(ω) – x)”</td>
<td>bₙ(1-x)</td>
<td>-bₙx</td>
<td>bₙ(1-u)</td>
<td></td>
</tr>
<tr>
<td>“bₙ(H(ω) – u)”</td>
<td>bₙ(1–u)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

¹⁸ Note that there would be no psychological difficulty in following this principle since it would require only the ability to recognize the syntactic difference between “H” and “P(H)=x)” or the verbal difference between “H” and “I believe that H”.

19
There is no world in which the subject has more than one set of odds on H, but here it is not because the subject’s value for \( P(H/H=x) \) is always whatever the 10th character in that expression is, but because the subject’s value for that expression is always whatever her value for \( P(H) \) happens to be. In worlds one and two that is x. In worlds three and four that is u.

The reason the 10th Character Principle works for those subjects for whom it works is that these subjects never discharge the condition \( P(H)=x \), i.e., become certain that \( P(H)=x \), unless \( P(H) \) does equal x. RSR is thus a special case of NGI. Following the 10th Character Principle will make anyone who does not have accurate beliefs about her beliefs vulnerable to sure loss, but the imperfect subject can avoid this entirely, regardless of the degree of her ignorance about her belief state, by following NGI. Since RSR is a special case of NGI, no subject will be vulnerable to sure loss from her value for \( P(H/H=x) \) if she follows NGI.  

Since NGI imposes an irrelevance between first- and second-order it may seem like a new way of making second-order probabilities idle, to put alongside the way that ST appears to do so by making their values extreme, and so probabilistically irrelevant to all other propositions. However, importantly, NGI only affects the subject’s responses to statements of her degrees of belief when taken by themselves. NGI does not imply \( P(H/H=x.S) = P(H) \), for S any statement of probability, which I will call No Interference (NI). I do not endorse the latter principle.

---

19 That RSR is a special case of NGI can also be seen by total probability:

\[
P(H) = P(H/P(H)=x)P(P(H)=x) + P(H/P(H)\neq x)P(P(H)\neq x)
\]

Following RSR the first term on the right hand side becomes x:

\[
P(H) = xP(P(H)=x) + P(H/P(H)\neq x)P(P(H)\neq x)
\]

If \( P(P(H)=x) = 0 \), that is, the subject is certain that she has degree of belief x

\[
P(H) = x(1) + P(H/P(H)\neq x)(0)
\]

it must also be the case that

\[
P(H) = x
\]

That is, she must fulfill Accuracy. However, that leaves several terms in the equation unused. If the subject follows NGI, then \( P(H/P(H)=x) = P(H) \) for all x, so total probability becomes:

\[
P(H) = P(H)P(H=x) + P(H)P(P(H)\neq x)
\]

The subject following NGI may have Accuracy or not without any sure-loss vulnerability, provided her confidence that she does not have degree of belief x equals 1 minus her confidence that she does.

\[
P(H) = P(H)c + P(H)(1-c)
\]
because a conjunct S in the condition could conceivably make the consideration that the subject believes H to degree x relevant to H. See Roush 2009, where an S stating that the subject is reliable to this or that degree about H is argued to make the proposition that she has a given degree of belief in H probabilistically relevant to H. Notably, the current sure-loss style argument does not defend \( P(H/P(H) = x, S) = x \) even if it were to work in defending RSR.\(^{20}\) Denial of NI also makes it possible for even extreme values of \( P(P(H) = x) \) to trigger change in the first-order probability of H, so second-order probabilities are not idle even for the self-transparent subject.

Defending failures of ST as rational allows it to be rational for someone to utter the Moorean sentence (both ommissive and commissive) that is, to both be confident that p and confident that she is not confident of p or not confident that she is confident. This does not imply that the Moorean false step is not a special kind of violation, but only that probabilistic coherence does not tell us what is special about it. Perhaps this should not surprise us since coherence is a generalization of deductive consistency, and most of us think that deductive consistency does not tell us what is wrong with Moorean sentences either.

4. The Broader Context: Better wrong than sorry

As we saw in response to the Direct Argument, being uncertain or wrong about what your degree of belief is does not yield vulnerability to sure loss in the classic sense, because of those possible worlds in which your degree of belief is what you actually think it is. Because of their existence you might have won on that bet, even though you actually did not. But that second-order bet does not tell the whole story of your wins and losses, since you will have actual win or loss, and possible wins and losses, from the first-order bet itself, and your first-order bet may leave you better or worse off overall than your second-order bet alone does. For example, the odds you actually have on the first-order proposition, say H, but do not, or do not fully, think you have on H, could leave you much better off on the question of H than the odds that you think you have on H would. And the absolute value of a win on H could be greater than the absolute value of the greatest loss you could sustain by being uncertain or wrong about what your degree of belief in H is. Obviously the reverse could also happen. More generally, the stakes on what the belief state is could be much lower than the stakes on H. We saw above that that is the way we would want it to be in an experiment to determine a subject’s degrees of belief.

\(^{20}\) NGI might seem to support the “bypass” view of knowledge of our beliefs (Hernandez 2013, Ch. 2), according to which the reason we consider p, and successfully use our reasons for believing p to determine whether we believe it, when we are asked whether we believe p is that we look past the proposition that we do or do not believe p to our grounds for believing p. But NGI asks to what degree we believe H given that we believe it to some degree, which is the reverse direction. Moreover, since relevance and irrelevance are symmetric the independence expressed in NGI implies \( P(P(H) = x/H) = P(P(H) = x) \), which says that H alone is irrelevant to whether I should believe that I believe H to some degree. This seems to conflict with the bypass view, although the exact correspondence between the two representations and claims is not clear.
The situation where the stakes are higher on the first-order proposition than on the question about your beliefs has a natural illustration in the context of implicit bias. Here, information being equal, the unbiased beliefs some of us think we have would lead to more accurate beliefs about who is most competent than prejudiced beliefs would, and so better outcomes for all in hiring. However, those may not be the beliefs we actually have, and for many of us, being wrong about what our beliefs are, whether racist or not, is just not as bad, emotionally, morally or for the self-interest of one’s organization, as being wrong about who is the most competent person to hire. We would say that you should accept the possibility that you have, say, racist beliefs – failing to know your own mind does not make you irrational – and focus on doing what you can to change the behavior that corresponds to such beliefs or the consequences of it, of not hiring the most competent applicant.\(^{21}\) It is better to lose the bet about what your beliefs are, and win at the first-order where those beliefs count.

\[21\] These remarks are elliptical for a complicated decision tree. If you are certain that you are unbiased, then you believe you have no reason to take steps to ameliorate the effects of prejudice on decision-making. Suppose that means you don’t. If you are right that you are unbiased, then you will win the bet on your beliefs and hire the most competent person you could have, other things equal: win-win. If you are certain that you are unbiased and wrong, then you will lose the bet on your belief, and, having been certain of your lack of bias, and having thus done nothing to ameliorate the effects of prejudice on decision-making, you will also decrease your chances of hiring the most competent person: lose-lose. If you are uncertain of your lack of bias, and you are biased, then you are right about your beliefs and, having taken ameliorative steps, you will also do as well as you could have at hiring the most competent person: win-win. If you are uncertain that you are unbiased, and you are unbiased, you will lose something in the bet on your belief, but you will take ameliorative steps, and do as well as you could have in hiring the most competent person: lose-win. The stakes on the second-order bet are lower, so unless ameliorative steps (e.g., blinding) are more costly than not hiring the most competent person, you are better off not being certain that you are unbiased.


