Sherrilyn Roush
Philosophy and Phenomenological Research 93 (3): 45-69, 2016.

Knowledge of Our Own Beliefs[1]

There is a widespread view that in order to be rational we must mostly know what we believe. In the probabilistic tradition this is defended by arguments that a person who failed to have this knowledge would be vulnerable to sure loss, or probabilistically incoherent. I argue that even gross failure to know one's own beliefs need not expose one to sure loss, and does not if we follow a generalization of the standard bridge principle between first-order and second-order beliefs. This makes it possible for a subject to use probabilistic decision theory to manage in a rational way cases of potential failure of this self-knowledge, as we find in implicit bias. Through such cases I argue that it is possible for uncertainty about what our beliefs are to be not only rationally permissible but advantageous.

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. 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, which the word "incoherence"[2] will refer to here, 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. Mistaken belief about one's own belief is a failure, but it is a lack of knowledge and not a failure of rationality in the sense expressed by the probability axioms.

[1] Thanks to audiences at Carnegie Mellon University, the Rutgers Epistemology Conference, the Berkeley-London Graduate Philosophy Conference, the London School of Economics, and King's College London for helpful criticism and discussion. Thanks in particular to Teddy Seidenfeld, Glen Shafer, Philip Dawid, Nick Shea, and Matt Parrott.

[2] Usage of this word varies between vulnerability to sure loss and violation of the axioms, two concepts that are largely extensionally equivalent but that can come apart depending on how sure loss is defined. One can violate the axioms and, arguably, not have the relevant vulnerability (Hacking 1967), and here the issue will be whether having not violated the axioms you can be vulnerable to sure loss by not having your higher- and lower- order beliefs in sync.

If a rational being needs reasonably good knowledge of her own beliefs, we will need more than the constraint of probabilistic coherence to explain why.

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. One might think that of all frameworks the probabilistic system should be able to tell us how to manage this situation because it is designed to tell us how to be rational in circumstances of incomplete information and uncertainty. But if probabilistic rationality requires perfect knowledge of our beliefs then there is no rational way to manage violation of the requirement. Thus in arguing that coherence does not require self-knowledge of belief, I make room for the welcome possibility that probabilistic decision theory can be brought to bear for weighing our options in light of possible ignorance of our beliefs. My arguments show more specifically how to use that framework, by identifying 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. We will see further that uncertainty about what our beliefs are is not only rationally permissible, but also can be advantageous, as for example in cases of potential implicit bias.

1. Is ignorance of our belief states even possible?

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 state of the world and 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 automatically give me 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 expectations about 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, which I think it is, 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 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

[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 problematic it would be statistically normal.

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 depends of course on what beliefs are. 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? How could it have a feeling it could not feel? 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 did not sense it?

Like many probabilists, I follow Frank Ramsey (Ramsey 1926, Armendt 2006) in rejecting the definition of a degree of belief as 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.[4]

But asking oneself questions, or verbal behavior generally, would not be the only way to investigate what one's degree of belief is. More tangible 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

[4] There are several kinds of attempted analyses of the notion of degree of belief appropriate for probability. These, including Ramsey's, are outlined and their flaws discussed in Eriksson and Hajek (2007), who go on to argue convincingly both that flawed analyses can have explanatory value and that we have every right and reason to take the concept of degree of belief as an epistemological primitive.

from tangible 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 a good bit depending on whether we are right or wrong about who is most competent. So we act on the basis of what we really think.[5]

Though Leonard Savage also thought of degree of belief as a state of mind that manifested itself in extraverbal behavior, 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, 27-30) 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, 27-30)

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 the image 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 compels actual betting, and even though actual betting is often 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

[5] Verbal behavior is of course also action; it can put valuable things at risk, even tangible things, and that positive or negative utility figures in the calculation of the most advantageous behavior overall. Being taken by others as unbiased on the basis of one's testimony to that effect will often have positive utility and if that utility has a greater absolute value for the subject than the negative utility of not hiring the most competent person, then implicit, that is, disavowed, bias, could be advantageous overall.

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.

The Ramsey view makes it natural to classify implicit bias as a case of (degree of) belief, because what makes a belief a belief on this view is its potential to affect the subject's actions, and the view has no requirement that she have conscious access to or voluntary control over this disposition. A person who believes she does not have racist beliefs would naturally tend also to believe that she would never disadvantage a minority in her decision-making during hiring. On the Ramseyan view, if she is wrong about her disposition to act – as the empirical evidence suggests most of us are – then she is wrong about her belief. There is a tendency in the literature to identify implicit bias as alief (Gendler 2008), an automatic or habitual belief-like attitude, particularly one disconsonant with one's explicit avowals, but on this definition of alief there is no inconsistency in classifying it also as a Ramseyan belief, because the latter view has no requirement that the disposition to act be responsive to reasons or easily changed.[6] This is important because it allows for the possibility of using probabilistic decision theory to manage the difficult situation implicit bias puts us in. In that framework probabilities are beliefs, so if an alief is not a belief then the framework is not available for those cases.[7]

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

[6] Some cases of alief will qualify as incoherent when described as Ramseyan degrees of belief. If for example you are willing to bet the same amount of money on each of two glasses you have seen be filled with sugar and water, on their being safe to drink, but are reluctant to drink one of the glasses, and not the other, because of a label "sodium cyanide" that you affixed to it yourself, then you are disposed to bet with two different sets of odds on the same proposition. But it does not seem inappropriate to classify such an alief as irrational.

[7] A sticking point in representing an alief as a belief will be whether the state is an attitude toward a specific propositional content or its intentionality is more diffusely distributed. I tend to agree with Mandelbaum (2013) that the state must have propositional content if it is to do the explanatory work that has been expected of it.

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.[8] 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. A bet 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 if we adopt the Ramseyan view of belief. As noted above, a subject may or may not have introspective access to what a given degree of belief of hers is because it is possible to have difficulty becoming aware of one's own dispositions to act. 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.[9] In this the framework stands in marked contrast to philosophers who argue for metaknowledge of our beliefs on the basis of a claim that rational belief revision requires us to have knowledge of our beliefs (e.g., Davidson 1984, Williams 2004: 208, Shoemaker 1994: 281-286, 1996: 33-34).

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.[10] Whether higher-order probabilities are legitimate or

[8] 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 it every day, and we are startled at the idea 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 or go anywhere at all. For most people that would not be a life worth having, so the beliefs and utilities are in balance.

[9] Dispositions corresponding to a conditional probability are needed in order for the subject to carry out conditionalization, but those do not amount to beliefs about the support relation. First, a conditional probability is not the probability of a conditional (Lewis 1976), and second, in the standard axiomatization conditional probabilities are equal to ratios of unconditional probabilities and are not some further thing beyond these parts.

[10] See Skyrms (1980) for a survey of and response to their objections.

need to be reigned in, there is no difficulty representing them since the claim that the value of a probability function is x is itself a proposition and so can fall within the domain of one or more probability 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 can be 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 the question of this paper 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.

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 a claim 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 or even entrenched. Thus Ramsey's view of belief allows us to add a type of evidence that we might have about our beliefs, and that might even 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 typically, 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 q for which the subject has some level of credence x,

either $P(P(q) = x) = 1$ or $P(P(q) = x) = 0$ and Confidence

if $P(P(q) = x) = 1$ then $P(q) = x$, Accuracy

Confidence says that if one has some degree of belief x in q, then one is certain that one's degree of belief in q is x or certain that it isn't. Accuracy says that if one is certain that one's degree of belief in q is x then indeed it is x. I will call these two conditions together "Self-Transparency" (ST).[11, 12, 13]

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 $P(B) = z$. With stake S, her gains when B is true and false are:

[11] This is a minimal property of knowledge, in which the subject's degrees of belief merely correspond to the facts. It lacks the additional robustness that epistemologists generally recognize as necessary for knowledge, which might come from justifiedness, or reliability, or tracking, or virtue. The minimal notion is sufficient for the purpose here, since if having this property toward one's degrees of belief is not necessary for rationality then a strictly stronger property is not either.

[12] 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. (Cf. Williamson 2000.) Using the new principle introduced below, 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 per se a failure that compromises her rationality. I address the perfect properties because the sure loss argument looks strong enough to defend even them.

[13] 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. For some problems and solutions for self-referential probability functions see Caie (2013), and Campbell-Moore (2015a, 2015b).

B true	B false
S - z·S	- z·S

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 - z·S, 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) is 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 who violates Self-Transparency there is a set of bets 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 necessary truth. A subject will be susceptible 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 five, 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.[14] 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 here.

Substituting a concealed coin toss for her degree of belief in the sure-loss argument above brings out the invalidity 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

[14] Thanks to Joe Ramsey for suggesting this comparison.

that it came up heads when it did not, i.e., the two ways of being most extremely wrong about H. 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. Now 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 the existence of 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 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 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. 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 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 only 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 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.

In treating worlds in which the subject has a different degree of belief from her actual one as irrelevant, the Direct Argument is either inconsistent with the subjective interpretation of probability – because it is assuming that epistemic uncertainty is not sufficient to make a variable random – or else it is begging the question of whether the subject could be rational without knowing what her beliefs are – by assuming the subject has no uncertainty or error about what her degree of belief is.

One might try replacing the Direct Argument with a denial that P(q) = x could even be an outcome or random variable, but one would need an argument for this 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 willing to stake something on a logical falsehood, but what justifies this is that the content of the proposition is a necessary truth.[15] No possible world in which the necessary truth is false is relevant because none exist. Since there are possible worlds in which the subject has a different degree of belief than the one she actually has, the burden is on one arguing for Self-Transparency to explain why we should treat her having a given degree of belief as necessary.

[15] 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.

It is a necessary truth relative to the subject's probability function, but resting an argument on that buys nothing. Her probability function as a whole could also have been otherwise.

Holding P(q) fixed when evaluating the subject's bet on q, and letting it vary when she is betting on the proposition P(q) = x will be enforced simply by treating the bets on q and the bets on P(q) = x as having different sets of relevant possible worlds. Evaluating the first-order and second-order parts of the subject's probability function thus differently may seem suspect since in evaluating coherence we want to know whether a person's degrees of belief fit together properly, and surely coherence of a whole function cannot be evaluated a piece at a time.

It is true that evaluating the coherence of a proper subset of a particular subject's degrees of belief is not enough to conclude that the subject is coherent, and evaluating each of an exhaustive set of subsets will not address whether they are coherent with each other. However, these are not what we are doing. Our question here is whether failure of transparency at the second-order about one's beliefs at the first order introduces incoherence, so we are entitled to assume that the first-order degrees of belief of the subject are coherent. In fact we must assume the subject is coherent at the first order in order to isolate our question. If we can see that inaccurate or uncertain second-order degrees of belief do not necessarily introduce incoherence into an otherwise coherent subject, then we will have certified the possibility of coherence of such a subject's entire set of probabilities, not just a subset of them, and not just subsets of them piecemeal but as a whole.

With this understanding, we are imagining a subject betting with the second-order part of her function on what the values of the first-order part of her function are, 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) ≠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) ≠ x, the subject gains z > 0. If her bet against P(q) = x corresponded to a degree of belief in P(q) ≠ x that was greater than 0, then she wins something. That non-actual world in which P(q) ≠ x and she wins something for betting even a penny in its favor is among the possible worlds relevant to the evaluation. An analogous point holds for the failure of Accuracy.

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 contingent beliefs are true or probable. This is why it makes sense at all 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

$P(q/P(q) = x) = x$ Self-Respect (SR)

called "Self-Respect" by David 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 "10^{th}-Character Principle" since it requires on the right hand side a re-inscription of whatever is in the 10^{th}-character position, counting from the left, an answer that the subject can give without first investigating how the world is. SR says that your degree of belief in H given that your degree of belief in H is x, should be x. We could paraphrase it: that your degree of belief is x is not a reason for it to be some other value.

SR is not trivially true. For example it does not say that if your degree of belief in q is x then your degree of belief in q is x. It does not follow from the probability axioms alone because it is a bridge principle between the orders and the axioms are within-order constraints that only impose relations between orders at extreme probability values. Notice also that SR does not mention or depend on what the subject's degree of belief in q actually is. The conditional probability can be written out as a ratio:

$P(q/P(q) = x) \quad = \quad P(q.P(q) = x) \mid P(P(q)=x)$

in which the expression "P(q) = x" never occurs naked, but only as an argument of the function P. This means that the principle makes stipulations on the basis not of what your degree of belief in q actually is but on what you *think* it is. It says that your supposing that your degree of belief in q is x does not give grounds for it to be some other value.

Assuming coherence, SR implies the Accuracy condition above because

$P(P(q) = x)) = 1$ and
$P(q/P(q) = x) = x$

together imply

$P(q) = x$

Confidence and Accuracy together imply SR, as shown by Sobel (1987, 69-70) and more simply by Christensen (2007, 325-6). Notably, SR alone does not imply Confidence.[16] SR is the synchronic instance of Bas van Fraassen's (1984) Reflection Principle, so Reflection presupposes and implies Accuracy about one's current beliefs. The principle has been popular: Koons (1992, 23) takes a version of SR to be a 'virtually undeniable principle … of rationality', van Fraassen takes the synchronic instance of Reflection as "-- I should think, uncontroversial" (van Fraassen 1995, 19), and Vickers (2000, 160) refers to SR as "a well-known principle of epistemic logic".

SR 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 SR 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 version of SR as an axiom to construct a theory of higher-order probability. (Gaifman 1986)

On the other side, Self-Transparency makes all of the second-order unconditional probabilities equal to 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.[17] ST is thus a perfect shield from the perceived trouble of second-order probabilities since it acknowledges the existence of the application of functions to propositions about their values that is licensed by the probability representation, while (apparently) rendering any hierarchies

[16] Assume SR. 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 SR, even if successful, cannot serve as a full argument for ST.

[17] The Savage-Woodbury "collapse" argument (Savage 1972, 58-59) 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 SR. It depends only on the fact that total probability relates the first- and second-order thus: $P(H) = P(H/P(H)=x)P(P(H)=x)) + P(H/P(H)\neq x)P(P(H\neq x))$, which does not imply SR (or the weaker RSR defined below). The collapse argument does not assume or imply ST and is not intended to. Its target is the attempted use of higher-order probability to represent imprecise credence.

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 a counterexample to it (Skyrms 1980, 125). 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 SR 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 SR, 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 a generalization of SR 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

H: The coin lands heads.

G: P(H) = x

we suppose that

P(H/P(H)=x) = y for some y > x,

that is, that the subject violates SR, and

P(P(H)=x) = z > 0,

that is, that the subject allows it as at least possible that the condition P(H) = x is fulfilled.[18]

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.

[18] I represent the subject's beliefs as probabilities rather than merely as credences because I am assuming they conform to the axioms, and are coherent until some specific assumption about the subject makes them otherwise.

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(ω) is 1 if H is true, and 0 if H is false, and similarly for G(ω), which is 1 if P(H) = x and 0 if P(H) ≠ x.

Outcomes --->	H and G	-H and G	-G
$G(\omega)b_1(H(\omega) - y)$	$b_1(1-y)$	$-b_1y$	0
$b_2(G(\omega) - z)$	$b_2(1-z)$	$b_2(1-z)$	$-b_2z$
Total each outcome			

There are only three columns because in case G is false the bet on the coin 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 here 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 H. That commitment must be listed in the table along with the other odds, yielding:

Outcomes --->	H and P(H) = x	-H and P(H) = x	P(H) ≠ x
$b_3(H(\omega) - x)$	$b_3(1-x)$	$-b_3x$	-------
$G(\omega)b_1(H(\omega) - y)$	$b_1(1-y)$	$-b_1y$	0
$b_2(G(\omega) - z)$	$b_2(1-z)$	$b_2(1-z)$	$-b_2z$
Total each outcome	$b_3(1-x)+ b_1(1-y)+ b_2(1-z)$	$-b_3x+ -b_1y+ b_2(1-z)$	$-b_2z$

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 b_1, b_2, and b_3 respectively. This gives a negative payoff in every column, as calculated in the final row below.

Outcomes --->	H and P(H) = x	-H and P(H) = x	P(H) ≠ x
$b_3(H(\omega) - x)$	$b_3(1-x)$	$-b_3x$	-------
$G(\omega)b_1(H(\omega) - y)$	$b_1(1-y)$	$-b_1y$	0
$b_2(G(\omega) - z)$	$b_2(1-z)$	$b_2(1-z)$	$-b_2z$
Total each outcome	**x-y + (y-x)(1-z)/2**	**x-y+ (y-x)(1-z)/2**	**(x-y)z/2**

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 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.

In addition, though the subject does not have a commitment to odds of x on H in the worlds of the third and fourth columns, 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) is something other than 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 SR, 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 SR, 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 – so we will record

this catch-all likelihood explicitly in the table in order to make plain that as far as we know it can take any value. The odds corresponding to this conditional probability we represent in general as v:1-v.

With these implicit terms spelled out, the table for a subject adhering to SR looks as follows.

Outcomes --->	H and P(H) = x	-H and P(H) = x	H and P(H)= u ≠ x	-H and P(H)=u ≠ x
	$b_3(1-x)$	$-b_3x$	$b_3(1–u)$, etc.	$-b_3u$, etc.
Actual odds ↓				
$(1-G(\omega))b_4(H(\omega) – u)$	0	0	$b_4(1–u)$	$-b_4u$
$G(\omega)b_1(H(\omega) – x)$	$b_1(1-x)$	$-b_1x$	0	0
$b_2(G(\omega) – z)$	$b_2(1-z)$	$b_2(1-z)$	$- b_2z$	$- b_2z$
Total each outcome	$(b_3+ b_1)(1-x)+ b_2(1-z)$	$-(b_3 +b_1)x + b_2(1-z)$	$(b_3+ b_4)(1–u) – b_2z$	$-(b_3+ b_4)u – b_2z$

Because this subject is following SR, the variable v in the first-row, third-column outcome world has become u. In none of the worlds indicated by the columns does the subject have more than one set of odds on H, or anything else, so no coefficients will lead to loss in all of the possible worlds represented.

However SR does not protect us from all trouble. The following conditions are compatible with those that were used to define the SR table:

1. P(H) = x
2. P(P(H) = x)) = 1
3. P(PR(H/P(H) = x) = y) = 1, y ≠ x, and
4. P(P(H) = x)) = 1 $\Rightarrow$ P(PR(P(H) = x) = 1) =1

where “PR” designates an objective probability function. Under these conditions the outcomes row showing the properties defining each possible world would have four more columns, and most saliently here, in half of those columns where it now has “P(H) = x”, it would also have “PR(H) = y”. If so, then for the subject with $G(\omega)b_1(H(\omega) – x) = b_1(1-x)$, that is, for whom P(H/P(H) = x) = x, there exist possible worlds in which she is willing to bet on H at odds x:1-x despite the fact, in that world, that the objective probability of H is y. It is no irrationality to have one’s degree of belief fail to match the objective probability – that is mere ignorance – but this particular failure is one the subject could have avoided because 1-4 imply that she is also certain that the objective probability of H is y. I will call this situation *awkward*.

We will avoid awkward questions in the current context by imposing a proviso that will remove such cases from the scope of the principle, thus:

P(H/P(H)=x) = x *provided no statement or set of statements of probability (other than* P(H) = x*) for which* P *has a value is (possibly together with* P(H) = x*) probabilistically relevant to* H.[19,20]

Call this "Restricted Self-Respect" (RSR). (Cf. Roush 2009, 253) A subject following this principle is safe from book since SR is safe and this is SR over a restricted domain. It tells her to follow the 10th character when there are no relevant probability statements that she has values for that could imply that the properties of the outcome-worlds are different from or in tension with what the condition P(H) = x in the conditional probability P(H/P(H) = x) designates them to be.[21] Analogously to SR, RSR implies a version of the Accuracy property – Accuracy qualified by the proviso – so the shift to RSR does not undermine but only qualifies the Indirect Argument.

However this does not imply that following SR or RSR is the only way to avoid incoherence. The more explicit table allows us to see that there is another safe way for a subject to accept a value for P(H/P(H) = x). Imagine that she has a policy of regarding the mere claim that she has a particular degree of belief in H as irrelevant to whether H is true. Thus, whichever possible 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 the disposition toward H is in that world. She would be following a strategy of regarding a mere statement *about* what her belief in H is as irrelevant to what her belief in H should be.[22] This is expressed by the following alternative principle to RSR:

---

[19] Note that this refers to probability statements, such as "P(H) = x" or "PR(H/P(H) = x)" that themselves have probability values. The function P also assigns values to statements such as H, but those first-order statements need not be disqualified by the proviso. The probabilities for all first-order propositions relevant to H should be taken into account in the value we give to P(H/P(H)), just as the probabilities for all same-order propositions relevant to A should be taken into account when evaluating P(A/B).

[20] Note that the paradoxical type of proposition treated by Caie 2013, where H is equivalent to P(H) ≥ .5, is ruled out by this proviso, since the equivalence makes P(H) ≥ .5 relevant to H, and the subject has a value for P(P(H) ≥ .5).

[21] The counterexamples to SR given in Skyrms (1980), Christensen (2007, 2010), Roush (2009), and Lasonen-Aarnio (2015) all appeal to suppositions relevant to the relation of the subject's beliefs to the world – optimism, drug consumption, empirical psychological evidence of unreliability, angel communications. They can be represented in the form of conditions 1-4 because those suppositions must be matters the subject has degrees of belief in if she is to be expected to take them into account. They are thus not counterexamples to RSR. A principle that handles such cases coherently without awkwardness when P(H) = x and PR(H/P(H) = x) = y are in the condition of the conditional probability is Cal: P(q/(P(q) = x and PR(q/P(q)=x)=y)) = y. (Roush 2009) A principle that handles the cases where 1-4 are in the background is: RSR when the proviso is fulfilled and P(q/P(q)=x) = y when conditions 1-4 are in the background, and are the only relevant statements of probability in the background. Analogous principles can be formulated for NGI below. These points can be seen by recognizing that what I call awkwardness here is a violation of the Principal Principle. See Roush 2009, 252-7.

[22] Note that there would be no more psychological difficulty in following this principle than there would for RSR since it substitutes for the ability to count ten characters the equally basic ability to recognize the syntactic

P(H/P(H)=x) = P(H) *provided no statement or set of statements of probability (other than* P(H) = x*) for which* P *has a value is (possibly together with* P(H) = x*) probabilistically relevant to* H*.*

Call this "No Gratuitous Interference" (NGI), and the same principle without the proviso "No Interference" (NI). RSR, analogously to SR, can be justified 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 then I have (probabilistic) reasons for whatever degree of belief I do have in H, but *that* I have that degree of belief in H is not all by itself a reason, at least not in general. As above with RSR, the proviso captures the intuitive idea that the principle applies only when the claim that the subject has a given degree of belief alone is under consideration. NGI, like RSR, is not awkward, but a subject following NI can be awkward for the same reason that one following SR can be: statements of probability can in conjunction with other statements of probability be probabilistically relevant to the properties of the relevant outcome worlds.

That a subject following NGI is not vulnerable to sure loss can be seen by writing out the table for the answers she gives:

Outcomes --->	H and P(H) = x	-H and P(H) = x	H and P(H)= u ≠ x	-H and P(H)=u ≠ x
	"$b_3(H(\omega) - x)$" $b_3(1-x)$	"$b_3(H(\omega) - x)$" $-b_3x$	"$b_3(H(\omega) - u)$" $b_3(1 - u)$	"$b_3(H(\omega) - u)$" $-b_3u$
Actual odds ↓				
$(1-G(\omega))b_4(H(\omega) - P(H))$	0	0	$b_4(1-P(H))$	$-b_4P(H)$
$G(\omega)b_1(H(\omega) - P(H))$	$b_1(1-P(H))$	$-b_1P(H)$	0	0
$b_2(G(\omega) - z)$	$b_2(1-z)$	$b_2(1-z)$	$- b_2z$	$- b_2z$
Total each outcome	$b_3(1-x)+ b_1(1-x)+ b_2(1-z)$	$-b_3x+ -b_1x+ b_2(1-z)$	$(b_3+ b_4)(1-u) - b_2z$	$-(b_3 +b_4)u - b_2z$

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/P(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

difference between "H" and "P(H)=x" or the verbal difference between "H" and "I believe that H". Both RSR and NGI would also require the subject who applies them to recognize whether the proviso is fulfilled, but while that cannot be guaranteed, the matter will often not be mysterious. It is a question of whether you have degrees of beliefs about whether or not something is interfering with your belief's relationship to the truth.

for P(H) happens to be. In the worlds of the first two outcome columns that is x. In the worlds of the third and fourth outcome columns that is u.

The reason the 10th-Character Principle works for those subjects for whom it does preserve coherence – the Accurate subjects – 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.[23] Following the 10th-Character Principle will make anyone who does not actually 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. Any subject can avoid sure loss by following NGI, but, unlike SR and RSR, NGI does not imply any version of the Accuracy property. Thus the Indirect Argument for Accuracy fails.

Since NGI imposes an irrelevance between the first- and second-orders 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. Also, probabilistic irrelevance is the same thing as independence, and if I have any epistemic competence at all then my believing p is not independent of whether p is true or not; why should I treat myself so disrespectfully by following NGI? Both of these worries are addressed by the proviso. NGI only applies to the subject's responses to statements of her degrees of belief when taken by themselves, and if as per the proviso there are no probability statements she has beliefs about (probabilities for) except the bare statement that she has degree of belief x in H, then she does not attribute to her beliefs any relation to the world, incompetent or competent. Thus, following NGI does not amount to disrespecting oneself epistemically. NGI and RSR are both ways of remaining neutral about oneself. The difference is the way they handle the risk of inaccuracy about one's belief in H.

On the other hand, when there are probability statements besides P(H) = x that a subject has beliefs about, NGI does not apply, and nothing says that conditioning on a statement of your degree of belief in that kind of case could not change your degree of belief. The awkward

---

[23] 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)≠ x)P(P(H)≠ x)
Following RSR the first term on the right hand side becomes x:
P(H) = **x**P(P(H)=x) + P(H/P(H)≠ x)P(P(H)≠ x)
If P(P(H)=x) = 1, that is, the subject is certain that she has degree of belief x, then
P(H) = x**(1)** + P(H/P(H)≠ x)**(0)**
So 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 the initial total probability equation above becomes:
P(H) = **P(H)**P(P(H)=x) + **P(H)**P(P(H)≠ 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)

situation described above is a class of cases in which the further assumptions 1-4 *make* the statement of what the subject's degree of belief is relevant to the probability of H. Principles for handling such cases have values other than x on the right-hand side, and so can underwrite changes in first-order degrees of belief on the basis of beliefs about one's beliefs by conditionalization, and this is so even if one has an extreme degree of belief about P(H) = x. (Roush 2009) Thus NGI does not force second-order probabilities to be inert.

One will notice that granting failures of ST as rational allows it to be rational for someone to assert Moorean sentences (both ommissive and commissive) that is, to both be confident that p and confident that she is not confident of p or confident that p and not confident that she is confident. This does not imply that the Moorean false step is not a violation of rationality, but only that probabilistic coherence does not tell us what is wrong about it. This should not surprise us since coherence is a generalization of deductive consistency, and most of us do not think that deductive consistency tells us what is wrong with Moorean sentences either.

4. The Broader Context: the value of uncertainty about one's beliefs

The Direct Argument for Self-Transparency failed because of the relevance of possible worlds in which the subject wins her second-order bet by having in those worlds different first-order degrees of belief from those she actually has. But the subject does not actually have those degrees of belief, and her having them is only subjectively possible, possible for all she knows. How could they really matter? They can matter in ways that make uncertainty about one's beliefs not only rationally permissible but advantageous.

They can matter for a combination of reasons. First, that one possesses a degree of belief in q is a state of the world. Like any state of the world it can have a utility, positive or negative, and that utility can be independent of the truth value of q. For example it is often of positive utility to believe that one's spouse or partner is faithful, whether he or she is or not. Relationships can be easier and more rewarding if the parties are not suspicious of each other. For another example, it could be of positive utility for an advertiser for RJ Reynolds to believe that smoking does not cause cancer, since it might help him sleep better at night.

Second, the bet a subject makes on her belief in q is not the only bet she makes – often she also bets on q, for example – and, as I will illustrate, the utilities of her bets at all orders can be weighed together, on the assumption that their units are commensurable. Third, like any belief, a belief about one's own belief can be motivating. In particular, as we will see below, it can motivate us to take actions that affect the outcomes of our first-order decisions. Finally, though a degree of belief different from your actual one is merely possible for all you know, this epistemic possibility matters to your decisions since you cannot base your decisions on reality itself but only on what you believe about reality. The coin may have already come up

heads, but if you have not seen the coin or been told how it landed, then the possibility that it is tails should play a role in any of your decisions that depend on that outcome.

The case of implicit bias illustrates these points nicely. Suppose the proposition in question at the first order is that women are by that fact less competent, call this H. We can imagine betting on whether or not one's degree of belief in H is 0 (one lacks bias) or greater than 0 (one has bias).[24, 25] What are the stakes on the bet whether or not one is biased? We may feel worse about ourselves if it turns out we have this belief that women are inferior. We may be embarrassed and troubled. We may feel okay, maybe even proud, if we do not have this belief. The stakes have mainly to do with our self-regard, except to the extent that our having belief in H leads others to think we have this belief and they penalize or reward us in consequence. The stakes on H itself, whether or not women are by that fact less competent, may be much higher. H is probabilistically relevant to whether Ms. Y or Mr. X is more competent, and my degrees of belief about that will affect my vote on which of them gets a job offer. The stakes for me on which candidate is more competent can be high. A more competent colleague will contribute more to my organization, and so to my job satisfaction, and I might care intrinsically about hiring the most qualified person because I care about fairness.

This is a case where beliefs about our beliefs can motivate us to act, or not. For example, one could imagine that if one were uncertain whether one had some disposition toward H one might take measures, such as blinding of applications, to prevent a possible such inclination from affecting a hiring decision. And suppose that if one were certain that P(H) = 0, that is, certain that one were unbiased, then one would not take mitigating measures. After all, one would be certain that there was no reason to do so.

We can see in the following table how things would turn out for a subject with these options.

[24] Having or lacking bias would be more fully represented as gender being probabilistically relevant to one's judgments of the competence of individuals rather than as a categorical proposition, but self-knowledge or ignorance of that relevance would be more complicated to represent and unnecessary for the current point. I assume that belief in H leads to gender being probabilistically relevant to one's judgments of competence, and non-belief not.

[25] Of course, in implicit bias cases there may in addition be flat preferences for, say, men over women that do their work in decision-making independently of beliefs about competence. These could be added into the picture. However, to represent implicit bias as involving only a flat preference and no (biased) belief about competence is not as charitable as I have been since on my picture we allow that the person is making a choice of candidates on the basis of a judgment about competence – however skewed or unknown to her – rather than only on a gut preference.

Second-order options → First-order options ↓	P(P(H) = 0) = 1	P(P(H) = 0) < 1
P(H) = 0	(3, 1)	(2, -1)
P(H) > 0	(-2, -1)	(2, 1)

The table has two dimensions of "choice"[26] and every choice and pay-off belongs to you. Your options at the first order are your beliefs about women, and are listed in the leftmost column. You are either unbiased, as in the first row of outcomes, or biased, as in the second row. Your options at the second order are listed in the top row. They are to be certain that you are unbiased, as in the first column of outcomes, or uncertain that you are unbiased, as in the second column of outcomes.

For the payoffs, the numbers are arbitrary except for the relations of greater and less. In the first position in the ordered pair we have the payoff from the first-order options, the option you take between the two rows, and in the second position of the ordered pair we have the payoff at the second order, from the option you take between columns, whether to be certain or uncertain that you are unbiased. The stakes on the second-order bet, 1 and -1, are taken here to be lower than those on the first-order bet, 3 and -2. For one who cares more about self-regard and the regard and reaction of others to one's beliefs than about hiring the most competent person, that relation would be reversed.

In figuring out the payoffs, we are assuming that H is false – otherwise the belief would not be a bias. The truth value of H affects your pay-offs because having a bias will make it more likely you vote for a less competent job candidate than you might have. In the upper left outcome you are unbiased, so you do as well as you could have at choosing the most competent candidate, +3, and you were certain that you were unbiased so you win as much as you could have on the second order bet, +1; this is a win-win. Below that we have the outcome where you are biased, so you lose some in your judgment of the competence of job candidates, -2, and you are biased but certain that you are not, so you lose there too, -1: lose-lose.

In the right column where you are uncertain about your belief, that uncertainty about whether you are biased made you take mitigation measures. Suppose for simplicity that those measures were perfectly effective, so whether you are actually biased or not you do as well as you could have at the first-order, in those judgments of candidates. However those positive payoffs are smaller than the one on the left, +2 rather than +3, because mitigation measures are not free, and that reduces your payoff from successfully choosing the most competent candidate. In the

[26] Despite the language of choice we need not assume that a subject has voluntary control over her belief states in order to make use of these tables. They only compare the values or advantages of various combinations of states and make no commitment about how you get or got to them.

second-order bet the subject uncertain of whether she is biased loses something in case she is not biased, so she gets a -1 in the second position in the upper right block. But, in the lower right block the subject uncertain that she's unbiased wins 1 because indeed she is biased; this is a win-win.

The subject on the left, who is certain she is unbiased, wins more than in any other possible outcome *if* she's right that she lacks bias. However, she loses more than in any other outcome if she is wrong about that. If the subject on the right is unbiased then she wins less than she could have because of the cost of mitigation measures and of being wrong about her belief, but if she is biased then she still has a win-win. Thus for a person concerned to avoid the worst-case scenario it is better to be uncertain about her belief, as long as the first-order stakes are higher than the second, with the proviso that if mitigation measures are costly enough the advantage of uncertainty could be outweighed. The table looks different if H is true[27], but the point here is that it *can* be advantageous to be uncertain what one's degree of belief is, not that it is always advantageous.

Being uncertain of your degree of belief is a violation of the Confidence direction of Self-Transparency, and we have found not only, above, that there is no good argument in sight that lack of Confidence brings incoherence, but also, just now, that lack of Confidence can be advantageous. There may not be situations where lack of Accuracy is advantageous, but many of us do have inaccuracies and the fact, argued earlier, that lack of Accuracy about our beliefs does not necessarily bring incoherence, means that as long as we follow NGI we can mobilize the framework of probabilistic decision theory for guidance on how to manage the hidden corners of our minds, in the way that I have illustrated here for failures of Confidence. It is not probabilistically irrational to be ignorant about what lies hidden in us, but since this fact makes available the probabilistic framework for handling the uncertainty it would be irrational not to use it.

Armendt, Brad (2006), "Frank P. Ramsey (1903-1930)" in *A Companion to Analytic Philosophy*. Edited by A. P. Martinich and E. David Sosa. Routledge, eISBN: 9781405133463, Ch. 10, 139-147.

Bertrand, Marianne and Sendhil Mullainathan (2004) "Are Emily and Greg more Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review* (94), 991-1013.

[27] If H is true then the subject who is uncertain about her belief has the greatest potential for gain (a total of +3), but also the greatest for loss (-4), whereas the subject certain that she is unbiased will either lose one or gain 2. Thus, whether it is advantageous to be certain or uncertain that one is unbiased depends on whether H is true, a fact that may lead to paradox or underdetermination.

Brandom, Robert (1994), *Making it Explicit*: *Reasoning, Representing, and Discursive Commitment*. Cambridge, MA: Harvard University Press.

---------- (2000), *Articulating Reasons*: *An Introduction to Inferentialism*. Cambridge, MA: Harvard University Press.

Briggs, Rachael (2009), "Distorted Reflection," *Philosophical Review*, Vol. 118, No. 1, 59-85.

Caie, Michael (2013), "Rational probabilistic incoherence," *Philosophical Review*, 122(4):527–575.

Campbell-Moore, Catrin (2015a), How to Express Self-Referential Probability. A Kripkean Proposal. Review of Symbolic Logic, available on CJO2015. doi:10.1017/S1755020315000118.

---------- (2015b), "Rational Probabilistic Incoherence? A Reply to Michael Caie", *Philosophical Review* 124 (3):

Carruthers, Peter (2011), *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.

Christensen, David (2007), "Epistemic Self-Respect," *Proceedings of the Aristotelian Society*, *Vol. CVII*, *Part 3*, 319-337.

Davidson, Donald (1984), "Thought and Talk," in *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press, 155-170.

---------- (1987), "Knowing One's Own Mind," *Proceedings and Addresses of the American Philosophical Association* 60 (3): 441-458.

Eriksson, Lina and Alan Hajek (2007), "What are Degrees of Belief?" *Studia Logica* 86: 183-213.

Gaifman, Haim (1986), "A Theory of Higher-order Probabilities," in Halpern, J.Y., (ed.), *Theoretical Aspects of Reasoning about Knowledge*. Los Altos, CA: Morgan and Koufmen, 275-292.

Gendler, Tamar (2008), "Alief and Belief", *The Journal of Philosophy* 105, no. 10: 634–663.

Hacking, Ian (1967), "Slightly More Realistic Personal Probability," *Philosophy of Science* 34, No. 4: 311-325.

Koons, Robert (1992). *Paradoxes of Belief and Strategic Rationality*. Oxford: Oxford University Press.

Lasonen-Aarnio, Maria (2015), "I'm Onto Something! Learning about the world by learning what I think about it", *Analytic Philosophy*. In press.

Lewis, David (1976), "Probabilities of Conditionals and Conditional Probabilities", *The Philosophical Review* 85: 297-315

Mandelbaum, Eric (2013), "Against Alief", *Philosophical Studies* 165: 197-211.

Moran, Richard (2001), *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton and Oxford: Princeton University Press.

Ramsey, Frank P. (1926), "Truth and Probability", in Ramsey, 1931, *The Foundations of Mathematics and Other Logical Essays*, Ch. VII, pp. 156-198, edited by R.B. Braithwaite. London: Kegan, Paul, Trench, Trubner, and Co. New York: Harcourt, Brace, and Co.

Roush, Sherrilyn (2009), "Second-Guessing: A Self-Help Manual", *Episteme* 6.3: 251-268.

Ryle, Gilbert (1949). *The Concept of Mind*. London: Hutchinson.

Savage, Leonard J. (1972). *The Foundations of Statistics*. Second edition, revised. New York: Dover Publications, Inc.

Sellars, Wilfrid (1963), "Empiricism and the Philosophy of Mind," in *Science*, *Perception*, a*nd Reality*. London: Routledge and Kegan Paul.

Shoemaker, Sydney (1994), "Self-Knowledge and "Inner Sense": Lecture II: The Broad Perceptual Model", *Philosophy and Phenomenological Research* (54): 271-290.

---------- (1996), "On Knowing One's Own Mind," reprinted in *The First-Person Perspective and Other Essays*. Cambridge: Cambridge University Press, 25-49.

Skyrms, Brian (1980), "Higher-Order Degrees of Belief," in D.H. Mellor (ed.) *Prospects for Pragmatism: Essays in Honor of F.P. Ramsey*. Cambridge: Cambridge University Press.

Sobel, Jordan Howard 1987: 'Self-doubts and Dutch Strategies', Australasian Journal of Philosophy, 65, pp. 56–81. 69-71.

Uchii, Soshichi (1973), "Higher-Order Probabilities and Coherence", *Philosophy of Science* (40): 373-381.

van Fraassen, Bas (1984), "Belief and Will", *Journal of Philosophy* 81: 235-256.

---------- (1995), "Belief and the Problem of Ulysses and the Sirens", *Philosophical Studies*, 77, 7–37.

Vickers, John M. 2000: "I Believe It, But Soon I'll Not Believe It Any More: Scepticism, Empiricism, and Reflection", *Synthese*, 124, 155–74.

Williams, Michael (2004), "Is Knowledge a Natural Phenomenon?" in Schantz (ed.), *The Externalist Challenge*. Berlin and New York; De Gruyter, 193-210.

Williamson, Timothy (2000), *Knowledge and Its Limits*. Oxford: Oxford university Press.