New Rational Reflection and Internalism about Rationality

Maria Lasonen-Aarnio

To appear in *Oxford Studies in Epistemology* 5

Abstract

It is often maintained that there are interesting, systematic connections between a subject's perspective on what it is rational for her to do and what it is in fact rational for her to do. For instance, epistemologists often deem as irrational certain kinds of mismatch between a subject's doxastic states and her judgments about what doxastic states it is epistemically rational for her to be in. *Rational reflection* principles are often put forth as an attempt to implement such ideas in contexts of uncertainty about what credences or degrees of belief are rational. After outlining some problems with Old Rational Reflection, this paper discusses what seems like a well-motivated fix, New Rational Reflection. It is argued that an intuitive way of trying to motivate the principle fails, and that it faces counterexamples. To say the least, the principle imposes substantial and controversial constraints on the kinds of epistemic situations it is possible to be in. The paper also discusses a more general problem with any attempt to formulate a rational reflection principle, which is that such principles take seriously certain kinds of uncertainty about what is rational, but not others.

1. Doing what one rationally takes to be rational

Does rationality require that there be a match between one's opinions about what it would be rational to do in one's situation, and what one does in that situation? For instance, could it be rational to simultaneously judge that rationality requires one to ___ (*to believe a proposition p; to have a certain credence distribution P; to desire, intend, or be motivated to ϕ; to ϕ*, etc.), while failing to___ ? Or, could it be rational to judge that rationality forbids one to ___, while nevertheless going on to ___? Numerous views, appearing in different sub-fields of philosophy, affirm the existence of interesting connections between opinions, or at least rational opinions, about what is rational and what is in fact rational.

For instance, many ethicists have argued that someone who believes that a rational agent in her situation would be motivated to ϕ, while failing to be motivated to ϕ, fails to be rational.[1] To some it has seemed almost uncontroversial that something

[1] For instance, an assumption along these lines plays a pivotal role in Smith's defense of internalism in ethics, and, relatedly, his proposed solution of the so-called moral problem (see, for instance, Smith (1994: 65)). Arpaly (2000) notes that it is "almost a universal assumption in contemporary philosophy" that it is never rational to act against one's best judgment about what it would be best for one to do in a given situation. Such views are instances of the trend I am

similar holds for belief: believing that her fully rational self would believe *p* – or simply, that it is rational to believe *p* – while failing to believe *p*, is a paradigmatic case of irrationality.[2] Similar-sounding principles have been defended for other doxastic states. Consider, for instance, the principle that if a possible rational agent is certain exactly what degrees of belief she ought, rationally, to have, then she has those degrees of belief.[3] And cases where one judges that it would be *irrational* to __, while nevertheless going on to __, might seem like even more egregious violations of rationality. Many have argued that it is irrational to believe (or have high confidence in) *p*, while believing (or having high confidence) that it is not rational to believe *p*, or that one's evidence does not support *p*.[4] That is, it is irrational to be in a state of *epistemic akrasia*.[5]

The aforementioned views, or variants thereof, are often assumed as uncontroversial premises for further arguments. Defences often appeal to the intuition that there is something incoherent about a subject who fails to match her own states with those she takes to be rational. Such an agent fails to be *rational by her own lights*, and failing to be rational by one's own lights, it is urged, is a paradigm failure of rationality.[6] Some even argue that such incoherence involves blatantly inconsistent states.[7] Examples abound of subjects who fail to follow their own advice about what is rational, and who seem clearly irrational: subjects who believe that the evidence does not support astrology while nevertheless believing in astrology[8]; subjects who stick to their evaluations of a body of evidence even after coming to believe that those

---

sketching if judgments about what it would be best to do, or what one should do, are judgments about what it would be rational to do.

[2] See, for instance, Smith (1994: 178). Titelbaum (*forthcoming*) argues that a subject who makes *any* mistakes about the requirements of rationality – whether about how they apply to one's own situation or to someone else's situation – is irrational.

[3] Elga (2013) finds this principle, which he calls *Certain*, almost self-evident, and I suspect that a lot of epistemologists would agree.

[4] See, for instance, Chisholm (1989: 6), Scanlon (1998: 25), Elga (2005), Hazlett (2012).

[5] See Horowitz (*forthcoming*) and Greco (*forthcoming*). Though note that *epistemic akrasia* is sometimes characterized as the state one is in when one believes a conjunction along the lines of *p, but my evidence does not support p*, or *p, but it is not rational for me to believe p.*

[6] See, for instance, Smith (1994: 65; 178) and Elga (2005).

[7] Greco (*forthcoming*) argues that epistemic akrasia involves an inner conflict between the outputs of distinct belief-producing psychological systems, an ultimately, inconsistent beliefs. Hence, Greco denies that when an agent believes that it is not rational for her to believe *p*, her belief is about a different, epistemological subject matter than her belief in *p*. However, it is fair to say that this is a minority view. Most would deny that epistemic akrasia involves attitudes that are strictly speaking inconsistent.

[8] Elga (2005).

evaluations are flawed due to fatigue or some other cognitive impediment[9]; subjects who believe that flying is safe while believing that their plane will crash[10], etc.

In effect, many of the views mentioned above are committed to something stronger than merely denying the rationality of a mismatch between the states one takes to be rational and the states one is in. For they are committed to the view that certain beliefs – or at least certain rationally permissible or rationally required beliefs – about what states it is (or is not) rational for one to be in are factive. Here is an instance of the kind of view I have in mind: if I rationally believe that it is (or is *not*) rational for me to believe *p*, then it is (or is *not*) rational for me to believe *p*. It is impossible to hold false but rational beliefs about what it is rational for one to believe. But let me first take a step back. Consider the following wide-scope schema:

Rationality requires that, if I believe that rationality requires that I __ , then I __.

Assume that I believe that rationality requires that I __. The wide-scope schema requires that I avoid the mismatch, but the mismatch can be avoided by not believing that rationality requires that I __. Rationality does not require that I __. Hence, wide-scope requirements that are instances of the above schema are compatible with the possibility of holding false beliefs about what rationality requires. But I very much doubt whether resorting to such wide-scope requirements a way of avoiding commitment to the factivity of *rationally required* belief about what is rational.

Assume that rationality requires that I believe that rationality requires that I __. The above wide-scope schema, together with the K axiom, entails that rationality requires that I __.[11] The following reasoning strikes me as very convincing: According to the wide-scope requirement, rationality requires of me that *either* I __, or that I fail to believe that rationality requires that I __. But if rationality requires that I believe that rationality requires that I __, then the only way in which I can respect the wide-scope requirement is if I __. At least rationally required beliefs about what rationality requires come out as factive. Now, Broome (2013) rejects the K axiom, and he might even reject its present application on the following grounds. Assume that I fail to believe what rationality requires, failing to believe that rationality requires that I __. Then, rationality

[9] Horowitz (*forthcoming*)

[10] Greco (*forthcoming*).

[11] Letting "O*p*" stand for "it ought to be the case that *p*", the K axiom is the following distribution axiom: $O(p \rightarrow q) \rightarrow (Op \rightarrow Oq)$. In the present context we can read "Op" as "rationality requires that *p*".

does not require that I __.[12] I disagree, but the issue is neither here nor there, for the question was whether rationally required *belief* about what is rational is factive. If it is rationally required of me that I believe that rationality requires that I __, and I *do* in fact believe this, then the above kinds of counterexamples are irrelevant. It is difficult to see how a proponent of the wide-scope schema could block the conclusion that rationally required belief about what is rational is factive.[13]

It is not surprising, then, that numerous authors explicitly express their views as narrow-scope requirements of rationality, or as the view that certain beliefs about what it is rational for one to believe are factive.[14] Moreover, note that these authors don't take their views to follow from the general factivity of rationally permissible or rationally required belief or certainty. And therein lies the main challenge: if even rationally required belief, opinion, or certainty is not in general factive, why is it in the special cases under issue? The challenge can be sharpened by considering the fact that beliefs about whether it is rational to believe a proposition *p*, on the one hand, and beliefs about whether *p*, on the other, seem to be about distinct subject matters. The first is about an epistemic subject matter – namely, the rationality of certain beliefs. The second is about the matter of whether *p* – perhaps, for instance, the matter of whether it will rain today. In so far as what opinions it is rational to hold is a matter of what one's evidence supports, denying the rationality of opinions that are mismatched in the way under issue is to pose constraints on the kind of evidence one could have.[15]

I will refer to the loose trend sketched as *internalism about rationality*. *Internalism*, because all of the views mentioned affirm some connection between one's

---

[12] Broome (2013: 120) gives the following counterexample to the K axiom. Suppose that you enter a marathon. As a result, prudence requires you to exercise hard every day. Prudence also requires that, if you exercise hard every day, you eat heartily. But assume that you are lazy, failing to exercise hard, even though you could. Broome thinks that in this case it does not follow that prudence requires that you eat heartily.

[13] Now, someone like Broome might object that *all* requirements of rationality are wide-scope, and that rationality could never require such a thing. However, such a view seems to be in tension with the idea that one is rationally required to hold beliefs that reflect one's evidence: if one's evidence supports a proposition to a sufficiently high degree, then one is rationally required to believe that proposition.

[14] Here are some examples: Hazlett's (2012) *Undercutting principle* is a narrow-scope requirement. Greco (*forthcoming*) takes opponents of epistemic akrasia to be committed to the claim that justified beliefs about what one ought to believe are impossible ("For the claim that epistemic akrasia is always irrational is really just the claim that a particular sort of justified false belief – a justified false belief concerning what one ought to believe – is impossible"). Smith (1994: 148) defends the following narrow-scope principle "If an agent believes that she has a normative reason to ϕ, then she rationally should desire to ϕ", where, on Smith's view, believing that one has a normative reason to ϕ is just believing that one would desire to ϕ if one were fully rational.

[15] Greco (*forthcoming*) spells out this objection forcefully.

internal perspective concerning what is rational – that is, one's (rational) opinions about what is rational – and what is in fact rational. This is not to say that what is rational is fixed by one's internal perspective: even if, for instance, rationally believing that believing *p* is rational (irrational) entails that believing *p* is rational (irrational), it doesn't follow that whenever believing *p* is rational (irrational), one must hold some opinion about whether this is so. The entailment need not go both ways. The idea is just that when one *does* hold (rational) opinions about what is rational, these opinions constrain what is in fact rational. Though the view bears similarities to various internalisms, the usage is by no means standard. For instance, within ethics, 'internalism' is often used to refer to a view on which there is a connection between one's judgments about what it is right to do (or what there is a normative reason to do) and what one is *motivated* to do. And in epistemology 'internalism' often refers to a view on which epistemic rationality (or justification, or…) supervenes on one's internal states.

My main focus below will not be on the kinds of internalist views about rationality just mentioned. Rather, I want to discuss the possibility of extending the idea that there must be a match between opinions about what is rational and what is in fact rational to a context of uncertainty about what is rational – and in particular, a context of uncertainty about what credences or degrees of belief it is rational for one to have. I will assume that such uncertainty can be perfectly rational. Consider a rational subject, Sophie, who has a credence distribution over what credence distributions it would be rational for her to have, or which credence distributions best reflect her evidence. (Unless otherwise indicated, when speaking about which credence distribution would be rational, I will mean which distribution is *rationally required*.)[16] Perhaps, for instance, she is 1/3 confident in each of the following hypotheses about the rational credence that it will rain today: the rational credence is 0.3; the rational credence is 0.5; the rational credence is 0.7. What credence in rain would reflect or match Sophie's opinions about the rational credence in a way that respects the spirit of internalism about rationality?

What should Sophie's credence in rain be, *conditional* on 0.3 being the rational credence that it will rain today? Well, it seems that it should be 0.3! And what should her credence be, conditional on 0.5 being the rational credence? 0.5. And so on. In this sense, Sophie should treat the rational credence function as an *expert* function. Then, her

[16] For simplicity I will focus on cases in which a subject has a credence distribution over a partition of hypotheses about the rationally required credence distribution and hence, cases in which the subject is certain that there is a uniquely rational distribution. Considering such cases will, I am hoping, teach us lessons that can be extended to cases in which a subject is uncertain whether such uniqueness obtains, and where at least some of the hypotheses she considers state that there is such-and-such a range of rationally permissible credence distributions.

credence in rain should be the weighted average of the candidate rational credences, weighted by her confidence in their rationality. Her credence in rain ought to be $1/3 \times 0.3 + 1/3 \times 0.5 + 1/3 \times 0.7 = 0.5$. I will call the resulting principle, following David Christensen, *Rational Reflection*:

Rational Reflection

$P_t\,(h \mid p_{\text{RATIONAL},\,t}\,(h) = r) = r$[17]

Here P_t is the credence function of a rational subject. *Rational Reflection* entails that a rational subject's *expectations* about what the rational credences are match her credences in a very straightforward manner: her credence in a proposition *p* equals her expectation of the rational credence in *p*. A narrow-scope requirement could be derived from *Rational Reflection* in the way sketched above. Assume that Sophie has a *rationally required* credence distribution over hypotheses about the rational credence in *p*. Because this distribution is rationally required, giving it up is not an option. And because *Rational Reflection* itself states a rational requirement, it follows that Sophie is rationally required to assign to *p* the credence that equals her expectation of the rational credence in *p*. This cannot be taken to follow from rational expectations in general being factive, for they are not. Again, we have a special connection between (rational) opinions about what is rational and what is in fact rational. And again, a challenge arises: why, in the special case at hand, do we have an exception to the non-factivity of rational expectations?

Rational Reflection has recently come under a line of attack.[18] However, many still feel that there is something right about the principle, and that rational opinions about what opinions are rational have got to constrain what is rational in some way that the principle tries, imperfectly, to capture. I will argue against this: *Rational Reflection* fails, and we shouldn't expect there to be any fix. In section 2 I briefly look at the problem with *Rational Reflection*, and a natural-seeming revision that has been proposed by Adam Elga, resulting in *New Rational Reflection* (NRR). In section 3 I discuss a line of argument for NRR employing as a premise a principle stating that if a rational agent is certain what credences it is rational for her to have, then she has those credences (*Certain*). I also introduce some formal apparatus to facilitate seeing what is

[17] In denoting probability functions, I will adopt the convention of using lower-case letters for non-rigid designators. So, "$p_{\text{RATIONAL},\,t}\,(h) = r$" should be read as "the ideal, rational credence in *h* (at *t*) is *r*".

[18] Williamson (2011) spells out a counterexample to the principle. See also Christensen (2010) and Elga (2013).

problematic about the argument. I then argue against NRR more directly. In section 4 I introduce a simple class of formal frames in which the principle fails, as well as describing candidate counterexamples. In section 5 I say why the principle seems to compromise intuitive motivations prompting the search for rational reflection principles. In section 6 briefly discuss a puzzling consequence of reflection principles, namely, that if a subject who is uncertain what credences it is rational for her to have is to satisfy such principles, then it looks like she will sometimes be required to hold a credence while being certain that it is irrational. Section 7 draws together some conclusions.

2. The problem with Rational Reflection: knowing more than the expert

As initially plausible as it may seem, *Rational Reflection* faces serious problems. These have been forcefully spelled out by other authors, so I won't spend too much time on them.[19] What I say below will constitute an additional line of argument against the principle, for I sketch counterexamples with a different structure from those that have been recently discussed.[20] But before seeing what is wrong with the principle, it may be useful to look at a slightly different formulation of the basic idea behind *Rational Reflection*.

Consider a subject who has a credence distribution over a partition of hypotheses about which (entire) credence function is rational for her. As above, I am using lower-case letters to stand in for non-rigid designators, and upper-case letters for rigid designators. So, for instance, "P_t" and "P*" rigidly name credence functions. Claims like "$P^* = p_{RATIONAL,t}$" should be read as "P* is the rational, ideal credence function at *t*".[21] Then, let *Old Rational Reflection* be the following principle:

Old Rational Reflection (ORR)
$P_t\,(h \mid P^* = p_{RATIONAL,t}) = P^*(h)$[22]

If a subject is rational, then her conditional credence in any proposition *h*, conditional on some function P* being rational, ought to be whatever P* assigns to *h*.

---

[19] See Williamson (2011), Christensen (2010), Elga (2013).

[20] Williamson (2011), Christensen (2010), and Elga (2013).

[21] Also, I will assume that a subject who assigns a credence to the propositions expressed is capable of thinking about the relevant credence function P* in such a way that it is transparent to her exactly what credences P* assigns to various propositions. So, for instance, she is not merely capable of thinking about the function through descriptions like "the credence function my older sister has at the moment".

[22] This is how Elga (2013) formulates *Rational Reflection*.

For a simple way to see what is wrong with ORR, consider the following entailment of the principle:

$$P_t\,(P^* = p_{RATIONAL,t} \mid P^* = p_{RATIONAL,t}) = P^*\,(P^* = p_{RATIONAL,t})$$

By the probability calculus, $P_t\,(P^* = p_{RATIONAL,t} \mid P^* = p_{RATIONAL,t}) = 1$. The upshot is that any probability function P* that is, from the perspective of P_t, possibly the ideally rational function, must be certain that it is the rational function.[23] Hence, ORR rules out its ever being rational to think that it even *might* be rational to be uncertain of one's own rationality. Note also that as long as P_t assigns some non-zero credence to P_t itself being the rational credence function, it must be certain that it is the rational credence function. In other words, ORR requires any credence function that is not certain of its own irrationality to be certain of its own rationality! Such entailments are especially bad given that our search for rational reflection principles was propelled by the observation that it is not always rational to be certain about what the rational thing to do in one's situation is.

I think these observations provide ample reasons to reject ORR, and there is no need to reiterate the counterexamples that have been recently discussed. Those unconvinced by the observations made still need to explain away the seeming counterexamples.[24] But maybe there is a core motivation for rational reflection principles that can be retained, and that *Old Rational Reflection* simply fails to capture. Here is an idea. Assume that you learn that Sophie is an expert with a perfectly rational credence function P*, but you also know that you have information that Sophie lacks (Sophie's total evidence is a subset of your total evidence). Then, it seems clear that you ought not to simply adopt Sophie's credence function. Instead, perhaps you ought to adopt the function that results from conditionalizing Sophie's function on all the extra evidence you have.[25] But note that merely learning that Sophie is an expert might give you evidence that Sophie lacks, for she might not know that she is an expert. Elga argues that this motivates the following thought: your credences (at *t*), conditional on Sophie's credence function being the expert function (at *t*), should be Sophie's credences,

[23] Elga (2013) makes this observation.

[24] Williamson (2011), Christensen (2010), and Elga (2013) discuss a particular kind of counterexample to ORR involving "clock beliefs". What is essential to such counterexamples is that which credence it is rational to assign to propositions about the position of the minute hand of a clock seems to depend on what the minute had in fact points to. Then, information about what credences it is rational for one to have gives one information about the position of the minute hand of the clock.

[25] Cf. Elga (2013).

conditional on her credence function being the expert function (at *t*). Here, then, is what might seem like a natural way of revising ORR:

New Rational Reflection (NRR)

$P_t\,(h \mid P^* = p_{\text{RATIONAL},t}) = P^*\,(h \mid P^* = p_{\text{RATIONAL},t})$

NRR looks promising: it doesn't require thinking that all credence functions that might be rational are certain of their rationality, and it doesn't fall prey to the recently discussed counterexamples to ORR. Note that NRR is analogous to the fix to another expert principle, Lewis's *Principal Principle*, suggested by Ned Hall and Michael Thau.[26] Moreover, the motivations given by Elga for *New Rational Reflection* are strikingly similar to those given by Hall for the *New Principal Principle*. (Ultimately, I suspect that the *New Principal Principle* is susceptible to problems that are structurally similar to the problems encountered by NRR, to be spelled out below, but this is not the place to pursue these worries.)

I will first discuss an argument from a principle mentioned above, *Certain*, to NRR. I then argue against NRR more directly, spelling out counterexamples.

3. An argument for New Rational Reflection

The following principle might seem almost self-evidently correct:

> *Certain*
> Whenever a possible rational agent is certain exactly what degrees of belief she ought rationally have, she has those degrees of belief.[27]

Now, one might think, as Elga does, that not only does *Certain* motivate rational reflection principles in the sense that there had better be a way of generalizing the principle to situations involving uncertainty about what is rational, but that at least in certain special cases *Certain* in fact entails instances of NRR. Assume that as a result of acquiring new evidence, I become certain what I ought to believe right now. If I ought to end up respecting *Certain* as a result of conditionalizing on this evidence, then a constraint is imposed on my initial state of mind. If the constraint is that my initial credences respect NRR, then NRR can be derived as a special case. This would seem to lend at least some support to NRR. An argument along these lines is not Elga's main way of motivating NRR. But it will be instructive to see why, even in these special cases, one

[26] Hall (1994), Thau (1994).
[27] Elga (2013).

cannot derive instances of NRR. For at first sight, the argument has a fair bit of intuitive appeal.

Here is the argument.[28] Assume that at a time t_0 Chandra is certain that the rational updating procedure is conditionalization, and that he is about to conditionalize on information about what credence function is rational for him at t_0. Chandra is in fact rational, and has probability function P_0 at t_0. These conditions are to specify the special case in which *Certain* yields instances of NRR. At a slightly later time t_1 he learns that credence function P* was rational for him at t_0. Then, at t_1, he has probability function $P_1 = P_0(\text{-}|P^* = p_{\text{RATIONAL},t0})$, which is rational for him.[29] But Elga also thinks that it follows from the conditions stated that Chandra is certain, at t_1, that the function $P^*(\text{-}|P^* = p_{\text{RATIONAL},t0})$ is rational for him right now, at t_1. If this is right, then by *Certain*, this function *is* rational for him. And then,

$$P_0(\text{-}|P^* = p_{\text{RATIONAL},t0}) = P^*(\text{-}|P^* = p_{\text{RATIONAL},t0}).$$

This is an instance of NRR.

The argument proceeds in two steps: first, $P_0(\text{-}|P^* = p_{\text{RATIONAL},t0})$ is rational for Chandra at t_1, and second, at t_1 Chandra is certain that $P^*(\text{-}|P^* = p_{\text{RATIONAL},t0})$ is rational for him at t_1. But for Chandra to have such certainty, it looks like he would have to be certain, at time t_1, of both (a) and (b):

(a) $P^* = p_{\text{RATIONAL},t0}$

(b) All the evidence I have acquired since t_0 is that $P^* = p_{\text{RATIONAL},t0}$

After all, if Chandra was merely certain of (a), but not of (b), how could he be rationally certain that now, at t_1, the rational probability function for him is P* conditionalized on the information that $P^* = p_{\text{RATIONAL},t0}$? Let us assume for now that (a) and (b) are distinct pieces of information: in conditionalizing on the former Chandra isn't automatically conditionalizing on the latter (just what it would be for (a) and (b) not to be distinct pieces of information is discussed below). Indeed, in general, learning a proposition *p* doesn't entail certainty that one learnt exactly *p*.[30]

---

[28] See Elga (2013), who notes that the argument adapts an idea from Ross (2006: 277-299).

[29] As I am using notation here, $P_0(\text{-}|P^* = p_{\text{RATIONAL},t0})$ is P_0 conditionalized on $P^* = p_{\text{RATIONAL},t0}$.

[30] This isn't entailed even by the assumption that certainty is luminous, iterating in a trivial manner:

Perhaps the problem could be fixed by conceding that Chandra needs to learn *both* (a) and (b). But that won't help. For even if we have now saved the second step of the argument (that is, Chandra is certain that $P^*(\text{-}|P^* = p_{RATIONAL,t0})$ is rational for him at t_1), the first step faces a problem (that is, that $P_0(\text{-}|P^* = p_{RATIONAL,t0})$ is rational for Chandra at t_1). If Chandra has learnt both (a) and (b), and these are distinct pieces of information for him, then it is *false* that all the evidence he has acquired since t_0 is that $P^* = p_{RATIONAL,t0}$. Hence, (b) is a proposition that is falsified as soon as Chandra learns it. If Chandra has learnt (b) by t_1, then it is not the case that the rational credence function for him at t_1 is $P_0(\text{-}|P^* = p_{RATIONAL,t0})$. After all, $P^* = p_{RATIONAL,t0}$ is only *part* of the information he has acquired since t_0, but the above argument really did really on the assumption that *all* Chandra has learnt by t_1 is $P^* = p_{RATIONAL,t0}$. An instance of NRR cannot be generated.[31]

The problem does not essentially rest on the fact that learning (b) is self-falsifying. I do not want to deny that it can ever be possible to know that one has learnt exactly such-and-such up to the present moment. Perhaps, for instance, it is possible to learn a self-referential proposition along the following lines:

> (b*) All the evidence I have acquired since t_0 is that $P^* = p_{RATIONAL,t0}$ and this very proposition.[32]

But replacing (b) with (b*) still does not yield an instance of NRR. In so far as (a) and (b*) are genuinely distinct pieces of information for Chandra, in conditionalizing on (b*) he isn't just conditionalizing on the information that $P^* = p_{RATIONAL,t0}$. But then, the rational credence function for Chandra at t_1 is not $P_0(\text{-}|P^* = p_{RATIONAL,t0})$. Let p^* be the conjunction of (a) and (b*). Even if we still get the result that $P_0(\text{-}|p^*) = P^*(\text{-}|p^*)$, this is not an instance of NRR, nor, as far as I can see, does it entail one.

Perhaps the argument still works on the assumption (a) and (b) are not genuinely distinct pieces of information for Chandra. And perhaps there is a way of understanding the assumption that Chandra is certain, at t_0, that he is about to conditionalize on information about what credence function is rational for him at t_0, on which this is the case. In particular, assume that though it may be possible for rational

---

If $P(p) = 1$, then $P(P(p) = 1) = 1$.

One might have such iterated certainty in a proposition p, while being uncertain that p is exactly what one learnt at a given time.

[31] Assuming that Chandra is certain that $P^*(\text{-}|P^* = p_{RATIONAL,t0})$ is rational for him at t_1, what we get is the following: $P_0(\text{-}|P^* = p_{RATIONAL,t0}$ and all the evidence I have acquired since t_0 is that $P^* = p_{RATIONAL,t0}) = P^*(\text{-}|P^* = p_{RATIONAL,t0})$.

[32] Thanks to a referee for prompting me to discuss this suggestion.

agents to conditionalise on (i.e. learn) false propositions, Chandra is certain that he is not going to conditionalize on a falsehood about which credence function is rational for him at t_0. For any of the probability functions P that Chandra considers as possibly rational at t_0, he is certain (at t_0) that P is rational at t_0 if and only if at t_1 he learns exactly the proposition that P is rational at t_0. Then, in conditionalizing on proposition (a) above, isn't he automatically also conditionalizing on proposition (b) – and doesn't he then become certain that $P^*(-|P^* = p_{RATIONAL,t0})$ is rational for him at t_1? But even given these assumptions, there is a glitch in the argument. To spell out the problem, it will help to introduce some tools from epistemic logic that I will employ more below, and that will help make explicit the assumption that (a) and (b) are not genuinely distinct pieces of information. My approach will closely parallel that of Williamson (2011).

A probabilistic frame $<W, R, P_{PRIOR}>$ will consist, first, of a set W of situations or cases. These need not be thought of as entire worlds. It might be more useful to think of them as something like centered worlds or cases. R is a relation between members of W which we can informally think of as that of epistemic accessibility: *w*R*w*' just in case the subject in *w* cannot epistemically rule out being in *w*'. We might express this by saying that for all she knows, she is in *w*'. But more generally, we can simply maintain that *w*' is epistemically accessible from *w* just in case being in *w*' is compatible with the subject's evidence at *w*. Note that it is *not* being assumed that epistemic accessibility must be reflexive. The formal framework can accommodate a view on which false propositions can constitute evidence. At least in the finite case, a rational subject assigns some non-zero credence to a case just in case it is epistemically accessible for her. P_{PRIOR} is a prior probability distribution over subsets of W. I won't assume the distribution to be uniform, assigning to each member of W the same weight. Now, we will also need to assign probabilities to propositions at members of W. To get the probability function P_w at a world *w*, we conditionalise P_{PRIOR} on the set of worlds accessible from *w*. Informally, we can think of this set as one's total evidence at *w*. Note that given this framework, two worlds have the same probability function just in case they can access exactly the same worlds.

In effect, what this apparatus enables us to model is a view that assumes updating to happen by Bayesian conditionalization on evidence thought of as sets of worlds or cases.[33] Given the apparatus introduced, here is what the assumption that (a) and (b) are not distinct pieces of information for Chandra amounts to: the set of accessible worlds (worlds accessible for Chandra at t_0) at which P* is rational at t_0 is

[33] Since the above argument from *Certain* to instances or NRR assumes updating to happen by such conditionalization, employing the framework should not bring in any illicit assumptions.

exactly the set of accessible worlds at which Chandra learns just this proposition at t_1. But care must be taken in spelling out just what the relevant proposition is. Since we are thinking about propositions as sets of worlds, to avoid confusion it is best to just explicitly talk about such sets. Let the set of accessible worlds (accessible for Chandra at t_0) at which function P* is rational at t_0 be S. Then, the assumption is that S is identical to the set of accessible worlds at which Chandra conditionalizes exactly on set S at t_1. As long as this is the case, the kind of problem raised above dissipates: in conditionalizing on S, Chandra is simultaneously conditionalizing (a) and (b) (or (a) and (b*) above), for from Chandra's perspective, both propositions are identical to set S. But there is still a problem.

At t_1 Chandra learns that P* is rational at t_0. That is, he conditionalizes on set S. By the assumptions made, it is true at every world in S that P* is rational at t_0, and it is also true at every world in S that since t_0 Chandra has conditionalized exactly on set S. Then, let us grant that Chandra is certain that *right now* (at t_1) the following function is rational for him:

$$P^*(\text{-}|S).$$

But since S just *is* the proposition that P* is rational at t_0, doesn't this entail that Chandra is certain that *right now* (at t_1) the following function is rational for him?

$$P^*(\text{-}|P^* = p_{\text{RATIONAL},t0})$$

It would be too quick to draw this conclusion, for just which set the proposition that P* is rational at t_0 is identical to depends on one's perspective: from the perspective of P_0 the proposition is set S, but it does not follow that from the perspective of P*, the proposition is also set S. This follows only if we are operating with frames that satisfy the following condition: the set of worlds at which P* is rational at t_0 and that are accessible for Chandra (at t_0) is *identical to* the set of worlds at which P* is rational at t_0 and that are accessible for worlds at which P* is rational at t_0.[34] This condition

[34] To see how the argument fails if this assumption is not made, consider a frame in which there is a world w^* at which P* is rational at t_0, but that Chandra cannot access at t_0. Nevertheless, assume that (all) worlds at which P* is rational at t_0 can access w^*. Then, from the perspective of worlds at which P* is rational at t_0, the set of worlds identical to the proposition that P* = $p_{\text{RATIONAL},t0}$ will contain w^*. And then, $P^*(\text{-}|P^* = p_{\text{RATIONAL},t0}) \neq P^*(\text{-}|S)$.

guarantees the desired instance of NRR, without any reliance on *Certain*.[35] The condition would follow from the assumption that Chandra can only become certain of and conditionalize on true propositions (and hence, it must be the case that $P_0 = P^*$), but this wasn't supposed to be assumed by the above argument.[36]

Hence, even assuming that (a) and (b) are not distinct pieces of information for Chandra, the argument from *Certain* to special cases of NRR does not go through. Of course, one could still argue for NRR using the kind of condition on frames just sketched as a starting point. But that would be an altogether different argument. Let me now turn to frames in which NRR fails, and to direct counterexamples to NRR.

4. A simple frame and counterexamples to NRR

Let a proposition that *unifies* two probability functions P and P* be a proposition *p* such that the function that results from conditionalising P on *p* just *is* the function that results from conditionalising P* on *p*. In other words, conditionalising the two probability functions on *p* produces the same function. In effect, what *New Rational Reflection* says is that for any probability functions P and P*, the proposition that P* is rational unifies P and P* so long as both P and P* assign some non-zero probability to P* being rational (and note that this is required for the relevant conditional probabilities to be defined in the first place).

Using the tools from epistemic logic introduced above, I will first describe a simple model (or, more precisely, frame) in which NRR fails. I will then argue that there are cases with the structure of the frame described, cases that constitute counterexamples to NRR.

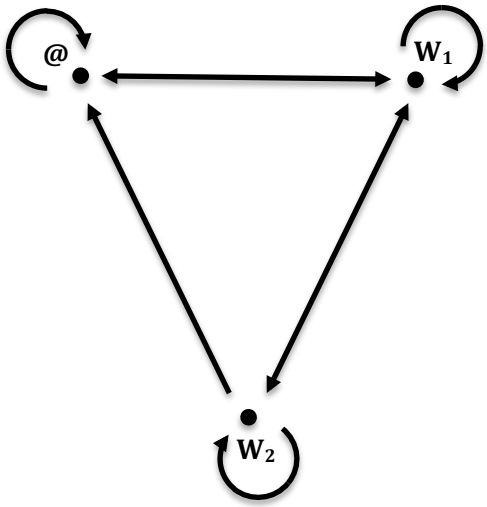
---

[35] Assume that the set of worlds at which P* is rational at t_0 and that are accessible for Chandra at t_0 (that is, accessible from worlds at which P_0 is rational at t_0) is *identical* to the set of worlds at which P* is rational at t_0 and that are accessible from worlds at which P* is rational at t_0. (Note that P* may or may not be identical to P_0). Let this set be S. Now consider $P_0(\text{-}|S)$ and $P^*(\text{-}|S)$. Let one's total evidence at worlds in which P_0 is rational at t_0 be E_0, and let one's total evidence at worlds in which P* is rational be E*. Then, $P_0(\text{-}|S) = P_{PRIOR}(\text{-}|E_0 \cap S)$, and $P^*(\text{-}|S) = P_{PRIOR}(\text{-}|E^* \cap S)$. Given the assumption that each world in S is accessible for both P_0 and P*, S is a subset of both E_0 and E*. But then, $P_{PRIOR}(\text{-}|E_0 \cap S) = P_{PRIOR}(\text{-}|S)$, and $P_{PRIOR}(\text{-}|E^* \cap S) = P_{PRIOR}(\text{-}|S)$ and hence, $P_0(\text{-}|S) = P^*(\text{-}|S)$. Because, from the perspective of both P_0 and P*, S just *is* the proposition that P* is rational at t_0, we get the following instance of NRR: $P_0(\text{-}|P^* = p_{RATIONAL,t0}) = P^*(\text{-}|P^* = p_{RATIONAL,t0})$.

[36] If Chandra only conditionalizes on truths, we can generate an instance of NRR by a much simpler argument. Assume that P_0 is rational for Chandra at t_0, and at t_1 he learns that P* is rational at t_0. Then, $P_0(\text{-}|P^* = p_{RATIONAL,t0})$ is rational for Chandra at t_1. Because Chandra only conditionalizes on truths, $P^* = P_0$. The resulting instance of NRR is trivial, and lends no support to NRR in its full generality.

A simple frame

Again, the frame <W, R, P_{PRIOR}> will consist of a set W of situations or cases, of a relation R between members of W which we can informally think of as that of epistemic accessibility, and of a prior probability distribution P_{PRIOR} over subsets of W. To get the probability function P_w at a world *w*, we conditionalise P_{PRIOR} on the set of worlds accessible from *w*. Informally, this set can be thought of as one's total evidence at *w*. Given this framework, two worlds have the same probability function just in case they can access exactly the same worlds. Here, then, is a simple, three-world frame in which NRR fails[37]:



W = $\{w_1, w_2, @\}$

R = $\{<w_1,w_1>, <w_2,w_2>, <@, @>, < w_1,w_2>, < w_1,@>, <w_2, w_1>, <w_2, @>, <@, w_1>\}$

W consists of three worlds or cases, @ being the actual one. w_1 and w_2 can access all cases, @ can access itself and w_1. Because w_1 and w_2 can access exactly the same worlds, they have the same probability function. Let P be the probability function (that is rational) at @, and P* the probability function (that is rational) at w_1 and w_2. Then, the proposition that P is rational is just $\{@\}$, and the proposition that P* is rational is $\{w_1, w_2\}$. To see that NRR fails for frames for which the above holds, consider the following:

$$P(\{w_1\} \mid P^* \text{ is rational}) = 1$$
$$P^*(\{w_1\} \mid P^* \text{ is rational}) < 1$$

[37] There is another kind of three-world frame as well, but I focus on this one, as it gives the structure of the kinds of examples described below.

This will hold as long as P_{PRIOR} assigns to each world a non-zero weight and hence, as long as each world assigns a non-zero weight to each world that it can access.

In a nutshell, the reason why NRR fails is that the set of worlds in which P* is rational that are accessible from @ is different from the set of worlds in which P* is rational that are accessible from the worlds in which P* is in fact rational. (Note that it was precisely this feature that also created trouble for the argument, discussed above, from *Certain* to instances of NRR.) As a result, the proposition that P* is rational is not a unifier of the two probability functions. Note that *Old Rational Reflection* (ORR) also fails in frames of the sort described, for:

$$P(\{w_1\} \mid P^* \text{ is rational}) = 1$$
$$P^*(\{w_1\}) < 1.$$

Those who are already convinced that epistemic accessibility is neither symmetric nor transitive should be surprised if we *couldn't* find examples with the structure of the simple frame described above. This – more so than the examples I describe below – is what ultimately convinces me that NRR has got to fail. By contrast, those already convinced that epistemic accessibility is transitive and symmetric will deny that the above frame is ever realized. But note that transitivity fails in the kinds of counterexamples to *Old Rational Reflection* recently discussed.[38] Still, a proponent of *New Rational Reflection* could argue, at the very outset, that at least symmetry holds. Now, I think that there are good, independent reasons to reject such symmetry: for instance, we can ordinarily rule out the possibility that we are brains in vats deceived about the world around us, but were we such brains in vats, we would not be able to rule out the possibility that we are in a good, ordinary case. But insofar as brains in vats and ordinary subjects can be in the very same internal states, those with internalist commitments about evidence will disagree. Below I discuss candidate examples of symmetry failure that cannot be dismissed just by appeal to internalism about evidence, for they don't involve pairs of cases in which one is in the very same internal states. I will also discuss a framework on which epistemic accessibility is symmetric for the reason that acquiring new evidence doesn't involve ruling out possibilities or worlds in the first place, but merely redistributing one's credences. As will become clear, insisting that updating happens by something like Jeffrey conditionalization will not, on its own, suffice to block counterexamples to NRR.

[38] Here I have in mind the "clock belief" types of cases described by Williamson (2011), Christensen (2010), and Elga (2013).

To start out with, assume the kind of framework discussed on which updating happens by Bayesian conditionalization on evidence thought of as sets of worlds or cases.

The ages of Agnes

We often estimate peoples' ages based on how they look. On several occasions I have been way off in my estimates. In particular, on several occasions I have presumed that a friend or acquaintance is in their 30's, later finding out, to my utter amazement, that they are in fact nearly 50. And on at least one of these occasions, I have at some point, after finding out the real age of my friend, seen a picture of them at 30, thinking "wow, she looks *much* younger in that picture – I can now see how her face is that of a 50-year old". By contrast, had I met only the younger, 30-year old version of my friend, I would never have thought that she might be 50. Hence, at least in some cases, and for some faces, the following asymmetry obtains: the younger face cannot easily be mistaken for a significantly older one, but the older face can be mistaken for a younger one. (I am not denying that this could sometimes go the other way, as with malnourished individuals, heroin addicts who later overcome their addiction, etc.) Consider, then, the following case.[39]

You are at the birthday party of your good friend Agnes, a sprightly centenarian. In order to mark the occasion, Agnes has hung up pictures of herself taken exactly 10 years apart, starting from birth, in each room of her house. You know that the floor you are in has three rooms, with pictures of Agnes as a 30-, 40-, and 50-year old. You are about to walk into a room, but don't know which. We can assume that you have not seen any pictures of Agnes at these ages (and you have only known Agnes since she was, say, 90), though you do know that Agnes has always been youthful, and that at 50 she was frequently mistaken for a 30-year old. Let @ be the case in which you step into a room and see a picture of Agnes at 30, w_1 the case involving a picture of Agnes at 40, and w_2 a case involving a picture of Agnes at 50.

You, in fact, walk into the room with a picture of Agnes at 30. Now, the following strikes me as perfectly plausible. You look at the picture, seeing the thick, black hair, porcelain-like skin, and hardly a wrinkle on her face: you know that this is not Agnes at 50. However, you cannot rule out her being 40 – after all, you know she never looked her age. At w_1 you see a 40-year old Agnes. Here you could go either way: you can't rule out her being 30, but neither can you rule out her being 50. At w_2 you see a picture of a

[39] Thanks to Ville Aarnio for dicussion.

50-year old Agnes. You see what looks like pretty thick black hair, a tanned face, and mountains in the background. There are some wrinkles around the eyes and mouth, but what you are looking at could be the face of a 30-year old who has spent lots of time outdoors, her face being exposed to the elements. This could be a picture of Agnes at 30, 40, or 50. You just cannot tell.

Instead of describing what you know in each of the three cases, I could simply have talked about which possibilities the evidence you gain by looking at the photographs allows you to rule out: when looking at the picture of a 30-year old Agnes, your evidence rules out the possibility that you are looking at a 50-year old Agnes, but not vice versa. All that really matters is that the view of evidence one is working with allows for the kind of asymmetry (and intransitivity) that the example relied on. Note that the kind of asymmetry described is compatible with internalist views of evidence, for one is not in the same internal states in any of the three cases described. And the precise numbers, of course, don't matter. What matters is just that the following holds: when looking at a picture of Agnes at a younger age of n years, you can (just) rule out her being some age $n + i$, whereas if you were instead looking at a picture of Agnes at $n + i$ years, you couldn't rule out her being n years.

Let P be the probability function at @. Since w_1 and w_2 can access exactly the same worlds (namely, all three), they have the same probability function. Let this function be P*. Now consider:

$$P(\{w_1\} \mid P^* \text{ is rational}) = 1 \neq P^*(\{w_1\} \mid P^* \text{ is rational}) < 1.$$

This is a failure of NRR. Before moving on to discuss an alternative framework, let me sketch a couple of recipes for generating more counterexamples to NRR.

Reals, fakes, and diminishing discriminative abilities

I take the following to be a pretty common phenomenon: coming across a fake, one can mistake it for the real thing, but when one sees (feels, hears, tastes, smells) the real thing, one can tell that it is not a fake. This is one reason why verbal directions are bad. It is easy to get lost by taking the little, barely visible path, where the guidebook tells one to take the *obvious* path to the left, and not to get distracted by tiny, barely visible paths. When seeing one of the small paths, it is far from obvious that it is not obvious – after all, it is a path. Upon finally finding the right path, one immediately recognizes that this is what the guidebook was talking about. This phenomenon can be exploited to construct more counterexamples to NRR. Assume, for instance, that the guidebook says that there

is a really obvious path that will lead to the lake, a slightly smaller but still clear path that will lead to the mountain, and, finally, a small, barely visible path that will lead nowhere. Looking at the obvious path, one knows that it is not the barely visible path. But one cannot rule out its being merely the slightly smaller path. Looking at the slightly smaller path, perhaps one cannot rule out its being either. And looking at the smallest path, one cannot rule out its being one of the larger paths – for it is clearly a path, and someone *could* refer to a path like that as really obvious.

Cases with the following feature provide resources for constructing yet further counterexamples to NRR: as we move along some parameter, our discriminative abilities get worse, and our margins of error grow. For instance, the more wine I drink, the worse I am telling how many glasses I have had.[40] The further I move from an object, the worse I am at telling how distant it is from me. The more works by Alexandre Dumas I read, the worse I am at telling how many I have already read. And as the minutes go by while I wait for my doctors' appointment, the worse I get at telling just how long I have waited. Consider the following example.

You are participating in an experiment. You know that you will be sat in a room for a period of 3 minutes (@), 5 minutes (w_1), or 7 minutes (w_2), after which a bell will ring, and the experiment will be over. During your time in the room, you will be asked to focus on a menial task preventing you from counting seconds in your head. The precise numbers don't matter, but it seems plausible that we could choose numbers in such a way as to generate a case with the sort of structure described above. The idea is that if you are in the case in which the bell sounds at 3 minutes ($w_@$), you know, upon hearing the bell, that 7 minutes have not passed. But you cannot rule out the possibility that 5 minutes have passed. If you are in a case in which 5 minutes have passed (w_1), you can rule out neither the possibility that only 3 have passed, nor the possibility that 7 have passed. And if you are in a case in which 7 minutes have passed, you can rule out neither 5 or just 3 minutes having passed, for your margin for error expands with time.[41]

While I am convinced by the above examples, some of the assumptions made could be resisted. Consider again *The Ages of Agnes*. You undergo different experiences at w_1 and w_2, since you are looking at different pictures of Agnes, taken 10 years apart. But then, wouldn't it be rational to have different credence functions at these two cases? Part of the objection may be resistance to the idea that you can only access three worlds.

---

[40] Thanks to Brian Weatherson for this example.

[41] Of course, the possibility of such examples depends on your margins for error growing in particular ways. While it seems plausible to me that the margins for error of at least some actual subjects do grow in the ways that generate examples with the structure described, all that is needed is that there are possible subjects for whom this is the case.

If more worlds are at play – as in more complex and perhaps more realistic examples – pairs of cases involving different experiences will typically be able to access different worlds. It is a challenge to make sure that we really have an example involving only three cases. But I suspect that the real source of the objection is resistance to the idea that updating on new evidence should always take the form of straight conditionalization: even if your evidence at w_1 or w_2 doesn't rule out any of the three cases, one might insist that some information is surely gained, information that should impact your probability functions at these worlds in different ways.[42] For instance, at w_1 you should be more confident that you are at w_1 than that you are at w_2, and vice versa. A related objection is that your evidence at @ doesn't allow completely ruling out w_2, even if you *know* that you are not at w_2. In particular, combining an internalist view of evidence with the idea that we nevertheless know a lot of the things we think we know leads to the though that a subject can know a proposition *p*, even if not-*p* is not ruled out by her evidence. If this is right then again, we cannot think of updating as happening by straight conditionalization on an evidence-proposition.

Let me now discuss the possibility of constructing counterexamples to NRR within a framework that assumes updating to happen by a more general form of conditionalization. As will become clear, such counterexamples can be resisted. But resisting them requires defending somewhat controversial assumptions.

Jeffrey conditionalization

As a starting point, assume that facts about a subject's evidence (and about which credence function is rational for her) supervene on her internal states, so long as it is not always perfectly luminous or transparent to a subject just what internal state she is in. So, for instance, even if a subject is in a different internal state in some case *w*' from the state she is actually in, she may not be able to rationally rule out being in *w*'. My diagnosis of *The Ages of Agnes* was at least compatible with such an internalist view of evidence.[43] Second, if evidence at least sometimes allows the ruling out of possibilities, it is not clear why such ruling out couldn't be asymmetric in a way that allows for the above kinds of counterexamples to arise. So let me discuss a view on which updating doesn't happen by ruling out worlds in the first place. We can no longer think of the probability function at a world as resulting from conditionalizing a prior probability

[42] Thanks to David Christensen for pressing this objection.

[43] I leave it open what counts as an internal state. Perhaps, for instance, all non-factive mental states are internal in this sense.

function on the set of accessible worlds, thought of as one's total evidence at that world. This means giving up the kind of formal framework sketched above.

Instead of Bayesian conditionalization, think of updating as happening by Jeffrey conditionalization. When one is at a world *w*, rather than there being an evidence proposition which is true in all the worlds accessible from *w*, there is a probability distribution over the worlds accessible from *w*, and we can think of the distribution as indicating what one's evidence is. Hence, being at a world is like undergoing a Jeffrey experience. Because straight conditionalization is a special case of Jeffrey conditionalization, the latter does, of course, allow the ruling out of worlds. But what I am envisaging is a view on which undergoing new experiences never makes it epistemically appropriate to completely rule out a possibility.

The counterexamples described above (case such as *The Ages of Agnes*) relied on different worlds accessing different sets of worlds. But counterexamples to NRR can also be constructed that rely not on accessing different worlds, but weighting the accessible worlds differently. Let S* be a set of worlds in which one is in some internal state E*, and let the probability function that is rational in all of these worlds be P*. If there is a probability function P (rational in some different set of worlds S in which, instead of E*, one is in some internal state E) that assigns different relative weights to the worlds in S* than P* itself does, NRR is in trouble. Assume that if one undergoes different experiences at two worlds *w* and *w*', it is not possible for the same probability function to be rational at these worlds. Then, counterexamples to NRR could be blocked by insisting that the partitions with respect to which Jeffrey updates happen never cut across experiential boundaries (in a way I discuss below). I will first discuss this constraint on partitions; I then say why the assumption that different experiences always mean different probability functions is implausible.

Consider a very simple example. As part of an experiment, you are sitting in pitch-dark room. When a bell chimes, there is a 2/3 chance that a sprig of lavender will placed right before you, and a 1/3 chance that there will be a sprig of mint. You also know that half of the sprigs of lavender have been genetically modified to smell of mint. Hence, just before the bell chimes, you consider the following three possibilities as equally likely: there will be a sprig of lavender that you can smell (@), a sprig of lavender that smells of mint (w_1), and a sprig of mint that smells of mint (w_2). The bell chimes, and you catch the distinct smell of lavender – hence, you are at @. How should you update your credences?

@	$\boldsymbol{w_1}$	$\boldsymbol{w_2}$
Smell lavender	*Smell mint*	*Smell mint*
Lavender	Lavender	Mint

Your experience doesn't allow you to completely rule out any of the three cases: though you are in perceptual contact with the sprig of lavender before you, you cannot be certain that it is lavender (rather than mint), or even that you are experiencing the olfactory sensation you seem to be experiencing (perhaps you are imagining things). There are four different partitions with respect to which you might perform a Jeffrey update: $\{\{@\}, \{w_1\}, \{w_2\}\}$ is the maximally fine-grained partition; $\{\{@, w_1\}, \{w_2\}\}$, which divides the worlds into those at which there is lavender before you and those at which there is mint; $\{\{@\},\{w_1, w_2\}\}$, which divides the worlds according to whether or not you experience the olfactory sensation as of lavender, and finally, there is $\{\{@, w_2\}, \{w_1\}\}$, the seemingly least natural partition, which divides according to whether the sprig before you is genetically modified.

Whether or not you end up respecting NRR depends on the partition you deploy for the purposes of a Jeffrey update. For instance, if, upon smelling the lavender at @, you are permitted to deploy the partition $\{\{@, w_1\}, \{w_2\}\}$ which divides worlds into those at which there is lavender and those at which there is mint, NRR is in trouble. Let the probability function that is rational at @ be P, and let the probability function that is rational at w_1 and w_2 be P*. First, how should P* weight w_1 and w_2? If the bell chimes and there is a smell of mint, you may not be able to rule out the possibility that you are catching a whiff of lavender after all, but you shouldn't think that it is likelier that there is lavender disguised as mint than that there is mint. That is, $P^*(w_1) \leq P^*(w_2)$. By contrast, consider your situation in @. You have a distinct olfactory experience as of lavender. You should raise your credence that there is a sprig of lavender before you. If you do a Jeffrey update by deploying the partition $\{\{@, w_1\}, \{w_2\}\}$, then because the probability of $\{@, w_1\}$ goes up from 2/3, that of $\{w_2\}$ goes down from 1/3, and @ and w_1 are equally probable (as they previously were), $P(w_1) > P(w_2)$. Then,

$$P(\{w_1\} \mid P^* \text{ is rational}) \neq P^*(\{w_1\} \mid P^* \text{ is rational}).$$

Again, NRR fails. The point is not restricted to the extremely simple model described. As long as one can at least sometimes deploy partitions that divide up worlds in accordance

with how things stand in the external world, it looks like opportunities for failures of NRR arise.[44]

If acceptable partitions always divide worlds according to the experiences one undergoes at those worlds, then the above reasoning is blocked. $\{\{@\},\{w_1, w_2\}\}$ is such a partition. If w_1 and w_2 are lumped together, then a Jeffrey update cannot change the ratios of the probabilities of these worlds. More generally, one might defend a principle according to which the ratios of probabilities of worlds in which one undergoes the same experiences, or in which one is in the same total internal states, are always preserved.[45] In defence of $\{\{@\},\{w_1, w_2\}\}$ – and against $\{\{@, w_1\}, \{w_2\}\}$ – one might complain that there is something very awkward about raising one's confidence in w_1 as a result of smelling lavender, even though one does not undergo any such experience w_1 (recall that in w_1 one has an olfactory experience as of mint, not lavender). But we shouldn't assume, without further argument, that smelling lavender couldn't directly give one information about whether there is a sprig of lavender before one even in the kind of example under consideration, without doing so via information about the experience one is undergoing.[46]

The question of what partitions ought to be deployed for the purposes of Jeffrey updates goes beyond the scope of this paper. But I very much doubt that the question can be resolved in a motivated way just by comparing intuitions. For instance, insisting on experience-partitions while denying scepticism seems to have consequences that are far from intuitive.[47] Assume that you open your eyes, and upon seeing your hands, have a perceptual experience as of hands. You are in normal conditions and there are no defeaters. The result we want is that it is rational for you to end up reasonably confident that you have hands. But assume that you perform a Jeffrey update by partitioning worlds according to which experience you are having. Given that you are undergoing a paradigm perceptual experience as of hands, it would seem rational for you to increase your confidence that you undergo such an experience (at least if, prior to opening your

---

[44] One objection to the kind of three-world model described is that if acquiring new evidence never justifies ruling out worlds, then one could never be in a model with only three worlds. But the point of using a three-world model was just to point to the kind of structural feature that allows for failures of NRR – namely, there being a set of worlds S* such that the probability function that is rational in those worlds weights the worlds (or subsets of worlds in S*) differently from some other probability function, rational in a different set of worlds.

[45] It is worth remarking again that such a principle would not suffice if the following was possible: the same credence function is rational at two worlds *w* and *w*', even though one undergoes *different* experiences at the two worlds.

[46] It is worth noting that Jeffrey himself didn't think that the relevant partitions should concern what experience one undergoes.

[47] The reasoning here relies on White's (2006) discussion of dogmatism. See Pryor (2013) for further discussion, including ways in which the reasoning could be resisted.

eyes, you were not already confident just what experience you would have). But if you increase your confidence that you are undergoing an experience as of hands, you will have to increase your confidence that you are a handless brain in a vat. Your new confidence that you have hands must be lower than your old confidence that you are not a handless brain in a vat.[48] To avoid scepticism, it looks as though the proponent of experience-partitions will have to concede that it is rational for you to be very confident *a priori* that you are not a handless brain in a vat. But this is far from intuitive.

It is worth re-iterating the commitments of a view that can hope to block counterexamples to NRR: Updating on new evidence takes the form of Jeffrey conditionalization, and one is never justified in completely ruling out a possibility. What one's evidence is at a world – and hence, which credence function is rational at that world – is fixed by one's internal state in that world. If one undergoes different experiences at two worlds, then different credence functions are rational at these worlds. Finally, the partitions one deploys in Jeffrey updates divide up worlds in accordance with the internal states one is in, or the experiences one undergoes, in those worlds.

Even if one granted that Jeffrey conditionalization should always happen with respect to experience partitions, the assumption that different experiences always make for different probability functions is far from uncontroversial. It is very plausible that what credences it is rational for a subject to have based on an experience she undergoes often depends on her discriminative abilities. But then, it should be possible for the same credences to be rational for two subjects, even if they undergo slightly different experiences. That overall experiential states fix rational credences does not entail that rational credences fix experiential states. Let w_1 and w_2 be two worlds in which one undergoes different experiences, but that have the same probability function. Then, even if Jeffrey conditionalization always happens with respect to experience-partitions, this would not be enough to guarantee that counterexample for NRR never arise, since the probability function that is rational at some third world w^* might weight w_1 and w_2 differently from the probability function that is rational at these worlds.

---

[48] Here is a sketch of the reasoning yielding this conclusion. Upon having an experience as of hands, you do a Jeffrey update by deploying a partition that concerns your experiences. Let S be the set of worlds at which you undergo an experience as of hands. Let *h* be the proposition that you have hands, and let *b* be the proposition that you are a handless brain in a vat undergoing an experience as of hands. *h* entails not-*b*. As a result of your experience, you increase your confidence in S, decreasing your confidence in at least some other members of the relevant partition. Each world at which *b* is true is a member of S – hence, *b* is a subset of S. But then, by increasing your confidence in S, you increase your confidence in *b*, and decrease your confidence in not-*b*. Because *h* entails not-*b*, the probability of *h* is at most the probability of not-*b*. Putting everything together, your new confidence in *h* is lower than you old confidence in not-*b*.

At first sight it didn't seem like the appeal of rational reflection principles rested on a specific way of thinking about evidence and updating. Just as the kind of synchronic constraint on rational credence functions posed by the Principal Principle seems to flow from the nature of chance, the kind of synchronic constraint on rational credence functions posed by rational reflection principles seemed, at first sight, to flow from the nature of what it is for something to be a rational credence function. But it turns out that *New Rational Reflection* rests on a somewhat controversial view about evidence and updating.

Let me now discuss a further set of problems for NRR, which is that it fails to respect internalism about rationality. The problems raised cast doubt on the whole project of trying to capture such internalism within a context of uncertainty about what is rational by means of some rational reflection principle.

5. Rational reflection and internalism about rationality

As we saw, NRR can be motivated by the rough idea that if I am rationally certain that Sophie is an expert (i.e. an ideally rational agent), but I have information that Sophie lacks, then I ought to defer to what Sophie thinks, *conditional on all the extra evidence I have*. I now want to discuss a further problem with NRR, which is that it doesn't respect the kind of internalism about rationality that was supposed to motivate rational reflection principles in the first place. Ironically, when I am uncertain just what my extra evidence is, the idea that I ought to adopt Sophie's credence function, conditionalized on all the extra evidence I have, goes against the rough idea that I ought to have the attitudes that I take a rational agent in my situation to have.

Assume that though I have extra evidence E (evidence that Sophie does not have), I have excellent reasons to believe that I have extra evidence E^*. Moreover, I know that while telling Sophie E would make her very confident in a proposition h, telling her E^* would make her very confident in not-h. Then, I have excellent reason to think that in my situation the perfectly rational agent Sophie would be confident in not-h. But if I ought to follow the rule of adopting the credence function that results from conditionalizing Sophie's credence function on the extra evidence I in fact have, then I ought to be confident that h is true. Hence, it is not at all clear whether NRR respects internalism about rationality within a context of uncertainty about what is rational in one's situation. Both *Old* and *New Rational Reflection* adhere to the internal perspective of a subject when it comes to the question of which credence function is rational, but not when it comes to the question of what the subject's extra evidence is.

This might prompt a search for a new rational reflection principle. Here is a thought. Assume that it is time *t*, and shortly thereafter at *t'* I learn (become certain) that P* is rational for me at *t*. What should my credence function at *t'* be? The problem was that $P^*(-|\ P^* = p_{RATIONAL,t})$ isn't necessarily the function I *take* to be rational for me, as I may not be certain (at *t'*) that what I have learnt between *t* and *t'* is that $P^* = p_{RATIONAL,t}$. Assume that at *t'* I am thus uncertain about my evidence, assigning a probability of 0.5 to having learnt *E* since *t*, and a probability of 0.5 to having learnt *E'*. Now, one might think that the following *would* respect internalism about rationality:

$$P_{t'}(-) = P^*(-|E) \times 0.5 + P^*(-|E') \times 0.5.$$

This suggests a more general thought, which can be heuristically expressed as follows. What should my credence in a proposition *h* be (at a time *t*), conditional on some probability function P* being rational for me (at *t*)? Well, what would I take myself to have just learnt were I to learn that P* is rational at *t*? My credence should just be P* conditionalised on that evidence. And what if, upon learning that P* is rational at *t*, I wouldn't be sure exactly what I just learnt? Well, then I should adopt the weighted sum of P*-conditional credences on the various pieces of evidence I think I might just have acquired, weighted by my credences that those are my new pieces of evidence. If $E_1, \ldots E_n$ are my candidates for the pieces of evidence acquired, and $r_1, \ldots, r_n$ are my credences, for each piece of evidence, that I just acquired it, then we can formulate the new principle as follows:

Internalist Rational Reflection (IRR)

$$P_t(-|P^* = p_{RATIONAL,t}) = P^*(-|E_1) \times r_1 + \ldots + P^*(-|E_n) \times r_n$$

But it is far from clear whether this rough idea is workable.

First, just what are $E_1, \ldots E_n$ and $r_1, \ldots, r_n$? I made use of counterfactual talk to express the gist of the new rational reflection principle IRR: "what would I take myself to have just learned, were I to learn that P* is rational at *t*?" It would be wise to clean up such talk, but there is no obvious way to do so. To get a glimpse of the problem of determining the values of $r_1, \ldots, r_n$, assume that there is some motivated way of deciding what $E_1, \ldots E_n$ are. Consider the following candidates for determining the weights $r_1, \ldots, r_n$:

(i) $r_i = P_t$ (I learn exactly E_i between *t* and *t'*)

(ii) r_i = P* (I learn exactly E_i between *t* and *t'*)
(iii) r_i = P* (I learn exactly E_i between *t* and *t'*| P* = $p_{RATIONAL,t}$)
(iv) r_i = P_t (I learn exactly E_i between *t* and *t'*| P* = $p_{RATIONAL,t}$)

(i) won't work, for the issue isn't what my opinions at *t* are concerning how likely it is that I will acquire various pieces of evidence at *t'*. What we are trying to get at (again, using the heuristic way of talking introduced above) is the opinions I would have were I to learn that P* is rational at *t*. (ii) is problematic for similar reasons. And conditionalizing probability function P* on the information that that P* is rational at *t* wouldn't seem to help, for we are interested in *my* perspective on what my evidence is, upon learning that P* is rational at *t*, not the perspective of P*. (iv) looks much better: instead of talking about the opinions I would have were I to learn that P* is rational at *t*, why not just talk of my present opinions, conditional on the proposition that that P* is rational at *t*? But within the present context, the problem with (iv) should be obvious: plugging it into IRR creates a regress, for precisely what we are trying to do is to come up with a rational reflection principle that tells us just which function P_t (- | P* = $p_{RATIONAL,t}$) is.

Second, I wonder whether even a principle along the lines of IRR is internalist enough. The above assumed that I am at least certain that conditionalization is the rational updating procedure: the strategy was to arrive at the probability function P_t (- | P* = $p_{RATIONAL,t}$) not by conditionalizing P* on what I would in fact learn were I to learn that P* is rational at *t*, but on what I would *take* myself to have learnt. But I may doubt not only whether I just learnt that P* is rational at *t*, but also whether I should take into account evidence by conditionalizing on – and if a lot of recent literature on higher-order defeat is on the right track, it may be rational for me to do so. Then, even if the sort of idea discussed above could be made to work, the resulting internalism would still seem half-baked.

One reply to such a worry is that in raising it I am bending the rules of the game, for it was simply being assumed that rational agents are certain of the rationality of conditionalization. But recall that the need for rational reflection principles arose within a context of uncertainty about just what is rational, and one way such uncertainty can arise is if one is uncertain about just what the right way of updating on one's evidence is. I see no motivated way of drawing a line between acceptable an unacceptable kinds of uncertainty. Before concluding, let me briefly discuss a further, somewhat related problem for rational reflection principles.

6. A potential conflict

In the very beginning I mentioned several views committed to the idea that certain kinds of mismatch between the states one takes to be rational and the states one is in are themselves irrational. It was noted that numerous epistemologists think that it is always irrational to be in a state of epistemic akrasia: one believes (or has high confidence in) a proposition *p*, while believing (or having high confidence) that it is irrational to believe *p*.[49] Consider now the state a subject is in if she assigns to *p* a credence or *r*, while being certain that it is irrational for her to assign to *p* a credence of *r*. Such a state is at least a close cousin of epistemic akrasia, and it might seem irrational for similar kinds of reasons:

> *Weak self-confidence*
> If a possible rational agent assigns credence *r* to *p*, then she is not certain that it is irrational to assign credence *r* to *p*.

This sounds like a pretty modest requirement – it doesn't even require rational agents to be moderately confident in the rationality of the opinions they hold. The requirement could be defended by a strategy mentioned at the very beginning. Consider a subject who assigns to the proposition that it will rain today a credence of 0.9, despite being certain that a credence of 0.9 is irrational. Such a subject holds a credence that is, by her own lights, certain to be irrational. Doesn't she exhibit a kind of incoherence that is rationally forbidden?[50] However, *Weak self-confidence* seems to conflict with the very idea of a rational reflection principle. Assume that Sophie is 0.5 confident that right now the rational credence to assign to the proposition that it will rain today is 0.3, and 0.5 confident that the rational credence is 0.7 (perhaps an epistemology oracle just told her so). The only way in which Sophie can respect *Old Rational Reflection* is by being 0.5 confident that it will rain. But because Sophie is certain that 0.5 is an irrational credence, by having that credence she would violate *Weak self-confidence*. And however ORR is revised, I see no reason why the new reflection principle would urge Sophie to assign to

[49] While many mean by 'epistemic akrasia' the state a subject is in when she believes a proposition *p*, while believing that it is irrational to believe *p*, there is no completely standard usage. For instance, Horowitz (*forthcoming*) means having high confidence in *p*, while having high confidence that one's evidence does not support *p*.

[50] I intend to set aside a subject who is certain that a credence of 0.9 in rain is irrational while having that credence, but who is confident that her credence in rain is, instead, some value other than 0.9. The issue could be avoided by adding to the antecedent of *Weak Self-Confidence* a clause specifying that the subject has sufficient epistemic access to her own credences.

p either a credence of 0.3 or 0.5 – in fact, such a recommendation would go against the spirit of rational reflection principles.

Further, it is at least unclear whether rational reflection principles are compatible with the view that it can never be rational to believe *p*, while believing that it is irrational for one to believe *p* – that is, the view that epistemic akrasia is always irrational. Assume a non-maximal threshold view of belief: to believe a proposition is to assign some credence above a value *r* to that proposition, where *r* is less than 1. Assume that the threshold for belief is 0.9, and that you know this. Assume that you have the following rational credences: your credence that the rational credence in *p* is 0.89 is 0.9, and your credence that the rational credence in *p* is 0.99 is 0.1. Then, your expectation of the rational credence is 0.9. If you respect *Old Rational Reflection*, your credence in *p* is 0.9. Given the 0.9 threshold for belief, you believe *p*. But you also believe that it is not rational to believe *p*. Hence, you are in the state of epistemic akrasia. Now, it is true that the reasoning relied on ORR. But there doesn't seem to be anything about NRR that would block the problem from arising – indeed, in many cases the two principles pose exactly the same constraint on one's credences.

Rational reflection principles sometimes recommend being in akrasia-like states. It is not clear whether this is a problem for rational reflection or the anti-akrasia principles discussed. But in any case, it does seem to put some pressure on the idea that rational reflections exemplify internalism about rationality within a finer-grained context of uncertainty about what is rational. Let me now draw together some conclusions.

7. Conclusions: giving up on rational reflection

My starting point was the rough idea that rationality requires that there be a match between one's opinions about what states it is rational to be in, on the one hand, and the states one is in, on the other. I dubbed the idea that (rational) opinions about what is rational have entailments for what is in fact rational *internalism about rationality*. I took the search for rational reflection principles to be spurred by an attempt to implement such internalism within a context of uncertainty about what is rational. Whatever one's favoured rational reflection principle is, any such principle seems to entail the following: if I have a rational credence distribution over hypotheses about which of P_1, ...P_n is the rational credence function for me, then that credence distribution fixes just which credence function is in fact rational for me. There is a very tight connection between one's opinions about what is rational and what is in fact rational.

While *Old Rational Reflection* faces serious problems, many still think that there is something right about rational reflection. Elga's *New Rational Reflection* (NRR) seems like a well-motivated fix that elegantly avoids the main problems and counterexamples for the old principle. The fix parallels a popular fix to another expert principle, Lewis' *Principal Principle*. I have argued against *New Rational Reflection*. First, one might hope to motivate the principle by showing that in certain special cases instances of it have to hold if one is to comply with a highly plausible principle, *Certain*. *Certain* denies that it could be rational to be certain that a credence function P is rational, while failing to have P. But, I argued, the argument from *Certain* to instances of NRR fails. I also argued that NRR faces counterexamples. Such counterexamples can be avoided, but this requires adopting a controversial view about evidence and updating. Finally, I discussed ways in which NRR fails to implement internalism about rationality – for while the principle defers to a subject's perspective when it comes to the question of what probability functions are rational for her, it does not defer to her perspective when it comes to the question of what her evidence is.

I think the above discussion warrants drawing at least some general lessons. First, NRR looks like a promising and well-motivated revision, but I hope to have shown that its viability rests on buying into somewhat controversial assumptions about evidence and updating. Such assumptions could be defended, but so could alternative ones. There is reason to think that the same will be true of any proposed reflection principle. However, the initial appeal of reflection principles seemed to rest on the nature of being a rational credence function. Just as it seemed like one ought to treat the chance function as an expert, it seemed that one ought to treat the rational credence function (or an agent with such a credence function) as an expert. What I am suggesting is that whether or not the rational credence function should be thought of as an expert function depends on very substantial epistemological issues.

Second, if the search for rational reflection principles is an attempt to implement internalism about rationality, there is reason to be sceptical about whether any such attempt *could* succeed. Any such principle will have to assume that certain facts about rationality cannot be rationally doubted: as we saw, even *Internalist Rational Reflection* cannot respect the internal perspective of a subject who doubts whether conditionalization is the rational updating procedure. Given that the need for rational reflection principles arose within a context of uncertainty about what is rational, this

raises the challenge of saying why we should only take certain kinds of uncertainty about what is rational seriously, and not others.[51]

Literature

Broome, John
(2013) *Rationality Through Reasoning*, Wiley Blackwell.

Chisholm, Roderick
(1989) *Theory of Knowledge*, third edition, Prentice-Hall.

Christensen, David
(2010) "Rational Reflection", *Philosophical Perspectives*, Vol 24.1: 121-140.

Elga, Adam
(2005) "On overrating oneself… and knowing it", *Philosophical Studies* 123: 115-124.
(2013) "The puzzle of the unmarked clock and the new rational reflection principle", forthcoming in *Philosophical Studies* 164.1: 127-139.

Greco, Daniel
(forthcoming) "A Puzzle About Epistemic Akrasia", *Philosophical Studies*, published online January 2013.

Hazlett, Allan
(2012) "Higher-Order Epistemic Attitudes and Intellectual Humility", *Episteme* 9.3:205-223.

Horowitz, Sophie
(*forthcoming*) "Epistemic Akrasia", *Nous*, published online May 2013.

Hall, Ned
(1994) "Correcting the guide to objective chance", *Mind*, New Series, Vol. 103, No. 412, pp. 505-517.

Lasonen-Aarnio, Maria
(*Forthcoming*) "Higher-order evidence and the limits of defeat", *Philosophy and Phenomenological Research*.

Pryor, Jim
(2013) "Problems for Credulism", *Seemings and Justification: New Essays on Dogmatism and Phenomenal Conservatism*, Chris Tucker (ed.), Oxford University Press, pp. 89-132.

Ross, Jacob
(2006) *Acceptance and practical reason*, PhD thesis, Rutgers University, October 2006.

Scanlon, Thomas

[51] I am extremely indebted to conversations with Ville Aarnio, David Christensen, Cian Dorr, Dmitri Gallow, Jim Joyce, Josh Schechter, Brian Weatherson, and Tim Williamson, as well as two *superb* referees for *Oxford Studies in Epistemology*. I also want to especially thank Alex Worsnip for very detailed and helpful comments on the paper.

(1998) *What we owe to each other*, Cambridge: Belknap Press.

Smith, Michael
(1994) *The Moral Problem*, Blackwell.

Thau, Michael
(1994) "Undermining and Admissibility" *Mind* 103: 491–504.

Titelbaum, Michael
(*Forthcoming*) "Rationality's Fixed Point (Or: In Defence of Right Reason)", *Oxford Studies in Epistemology*.

White, Roger
(2006) "Problems for Dogmatism", *Philosophical Studies* 131: 525-557.

Williamson, Timothy
(2011) "Improbable knowing", In T. Dougherty (ed.), *Evidentialism and its Discontents*, Oxford University Press.