Synthese (2016) 193:3099–3145
DOI 10.1007/s11229-015-0922-y

A non-probabilist principle of higher-order reasoning

William J. Talbott[1]

Received: 14 May 2014 / Accepted: 16 September 2015 / Published online: 9 October 2015

Abstract The author uses a series of examples to illustrate two versions of a new, nonprobabilist principle of epistemic rationality, the special and general versions of the metacognitive, expected relative frequency (MERF) principle. These are used to explain the rationality of revisions to an agent's degrees of confidence in propositions based on evidence of the reliability or unreliability of the cognitive processes responsible for them—especially reductions in confidence assignments to propositions antecedently regarded as certain—including certainty-reductions to instances of the law of excluded middle or the law of noncontradiction in logic or certainty-reductions to the certainties of probabilist epistemology. The author proposes special and general versions of the MERF principle and uses them to explain the examples, including the reasoning that would lead to thoroughgoing fallibilism—that is, to a state of being certain of nothing (not even the MERF principle itself). The author responds to the main defenses of probabilism: Dutch Book arguments, Joyce's potential accuracy defense, and the potential calibration defenses of Shimony and van Fraassen by showing that, even though they do not satisfy the probability axioms, degrees of belief that satisfy the MERF principle minimize expected inaccuracy in Joyce's sense; they can be externally calibrated in Shimony and van Fraassen's sense; and they can serve as a basis for rational betting, unlike probabilist degrees of belief, which, in many cases, human beings have no rational way of ascertaining. The author also uses the MERF principle to subsume the various epistemic akrasia principles in the literature. Finally, the author responds to Titelbaum's argument that epistemic akrasia principles require that we be certain of some epistemological beliefs, if we are rational.

✉ William J. Talbott
wtalbott@u.washington.edu; wtalbott@uw.edu

[1] Department of Philosophy, University of Washington, Box 353350, Seattle, WA 98195-3350, USA

Keywords Higher-order reasoning · Probabilism · Bayesianism · Dutch book · Calibration · Logical omniscience · Epistemic akrasia · Fallibilism

1 Introduction

This is a paper about reasoning with what is called *higher-order evidence* (e.g., Kelly 2010; Christensen 2010a)—or, more precisely, higher-order reasoning about the reliability of one's own degrees of belief. Agents who engage in this kind of reasoning are typically agents who make mistakes and can recognize that they make mistakes. They use that information to evaluate their own reliability. In the cases of interest, they also recognize that other agents have cognitive processes similar enough to theirs that information about the kinds of mistakes that those other agents make can be relevant in evaluating their own reliability. My goal will be to articulate metacognitive principles that explain how and why it is that evidence of this kind often provides a certain kind of *undercutting defeater*, a *reliability defeater* (Pollock 1984, pp. 112–113), for a rational agent's antecedent beliefs or degrees of belief.

To explain my principles, I will contrast them with the following kinds of normative theories, classified by the kinds of normative constraints they include:

(1) *Synchronic constraints* As I use the term, *probabilists* take some version of the probability laws as synchronic constraints on rational degrees of belief (Howson and Urbach 1989, pp. 16–17). Probabilists can be further categorized as follows: (1) *strict classical probabilists* (e.g., de Finetti [1937]) assume that *all* instances of *all* the laws of classical logic have probability of 1.0; (2) *non-strict classical probabilists* (e.g., Garber 1983) assume that *all* instances of *some* (but not all) of the laws of classical logic have probability 1.0 (typically the truth functional laws); (3) *strict non-classical probabilists* assign probability of 1.0 to *all* instances of *all* the logical laws (at least one of which is non-classical). Most of my discussion will focus on strict classical probabilism.

(2) *Diachronic constraints* As I use the term, *Bayesians* are classical probabilists (strict or non-strict) who take some version of conditionalization as a constraint on at least some empirical reasoning—that is, rational changes in degrees of belief based on empirical evidence. *Global* Bayesians hold that only empirical reasoning that satisfies some form of conditionalization is rational (or coherent). *Non-global* Bayesians hold that some form of conditionalization is a constraint on some empirical reasoning, but they allow for other kinds of empirical reasoning to be rational (or coherent), also.

Although my goal is to articulate a principle that applies to all cases of higher-order reasoning, I focus on cases that have bedeviled classical probabilists, and, thus, Bayesians; the problems apply to non-classical probabilists, also. I will use the examples to motivate consideration of two versions of a non-probabilist principle of empirical metacognitive reasoning.

This paper is not simply another listing of the idealizations of Bayesianism or non-Bayesian probabilism. The role of the examples is to provide clues to help us articulate new principles of higher-order reasoning that can explain them. In this paper I employ a bottom-up methodology in which I rely on judgments about particular cases—particular cases of rational and irrational synchronic degrees of belief and rational and irrational diachronic changes in degrees of belief—to provide the evidence for my two higher-order explanatory principles. Advocates of Bayesian or other probabilist views may be inclined to dismiss my examples as departures from what I will call *'idealized' rationality*, which is what they want a theory of. To signal this difference of approach, I will say that I am attempting to articulate a metacognitive principle of *human* rationality, not 'idealized' rationality. However, by the end of the paper, I will have cast doubt on whether the Bayesian or other probabilist models of 'idealized' rationality are a kind of ideal that humans should even aspire to.

Rather than keep it as a surprise, let me describe upfront the kinds of cases that won't have Bayesian or other probabilistic solutions. Call any form of reasoning that reduces a confidence assignment of 1.0 to less than 1.0 *certainty-reducing reasoning*. No form of conditionalization could be used to explain how certainty-reducing reasoning could be rational. All probabilist theories require that some propositions—at least some instances of logical truths and some epistemological truths—be certain. No probabilist could explain the rationality of certainty-reducing reasoning for those propositions.

Let a *fallibilist* assignment of degrees of belief be one that assigns no proposition degree of belief of 1.0. Clearly, no probabilist could explain how it could be rational to adopt a fallibilist assignment of degrees of belief, nor could they explain the reasoning that would lead to such a result. In this paper I articulate non-probabilist principles of metacognitive rationality that can even explain reasoning that leads to fallibilism.

To signal that my principles are non-probabilist, I say that my theory is a theory of rational *degrees of confidence*. Also, I do not assume that confidence assignments are *complete* in the sense that an agent assigns confidence to every proposition (or every proposition that she can conceive of). I allow for agents to assign confidence to some propositions and not others—depending, for example, on what propositions are of interest to them.

In the literature, the standard defenses of probabilism are arguments that degrees of confidence that fail to satisfy some version of the probability axioms are rationally defective in some way. Another role of the examples I discuss here is to help to indirectly undermine those arguments, by illustrating a kind of reasoning that is paradigmatically rational for human beings (even if the agents are regarded as non-idealized). I will also address the most influential of the arguments for probabilist synchronic constraints on rational confidence assignments to show that their most important insights can be translated into a non-probabilist framework. In particular, I will show that it is possible to endorse external calibration and accuracy maximization (inaccuracy minimization) as epistemic goals of rational degrees of confidence without requiring that they satisfy the probability axioms. My principles of metacognitive rationality will actually be special cases of what Jeffrey (1986) calls *de Finetti's law of small numbers*. Since I hold that calibration and accuracy maximization are goals of rational degrees of confidence and my own principles can be understood to be an implication of de Finetti's law of small numbers, it might seem that I myself am committed to some form of probabilism. I take up this objection below.

Finally, let me just say that, though I discuss a variety of probabilist views, I could not possibly discuss them all. The criticisms that I make often generalize to others that

I don't discuss. Also, for reasons of space, I don't have an opportunity to balance my criticisms of the views that I discuss with an appreciation of their strengths, of which there are many. There is simply not space to present a balanced evaluation of them here.

2 Examples

1. Fallibility about evidence. While working on the structure of DNA, Crick and Watson were greatly affected by the failure of their colleagues at the Cavendish Laboratory to discover the structure of one kind of polypeptide chain in protein before Linus Pauling made the discovery. Their Cavendish colleagues failed because they had tried to come up with a structure that explained all of the experimental evidence; Pauling just ignored one piece of evidence that did not fit his model. Here is Crick's description of his and Watson's response to that failure:

"Because of it, I argued that it was important not to place too much reliance on any single piece of experimental evidence . . . . Jim was a little more brash, stating that no good model ever accounted for all the facts, since some data was bound to be misleading if not plain wrong" (Crick 1988, pp. 59–60). *Suppose that before discovering their colleagues' mistake, Crick and Watson had regarded their evidence as certain. How could the empirical certainty-reducing reasoning that led them to reduce their confidence in their evidence be rational?*

This example will not, by itself, cast doubt on all probabilist views, but it will raise problems for Bayesian views that allow that it can be rational to be certain of empirical propositions, but rule out the possibility of rational certainty reductions on the basis of empirical reasoning. Consider the following simple form of conditionalization:

(*Simple Conditionalization*) Rational changes in degrees of confidence from prob_1 (at time t_1) to prob_2 (at a later time t_2) require that, where E is one's total evidence E acquired between t_1 and t_2: For any proposition P, $\text{prob}_2(P) = \text{prob}_1(P/E)$.

Let *simple, global Bayesianism* be the view that all rational changes in degrees of confidence due to empirical reasoning are instances of simple conditionalization. Simple, global Bayesian views require that evidence be certain, but they cannot explain how certainty reductions in one's evidence could be a rational form of empirical reasoning—that is, how it could be rational to change what is taken to be evidence on the basis of empirical reasoning. In fact, no form of conditionalization could explain the rationality of empirical certainty-reducing reasoning. So no global Bayesian could explain the rationality of such a change.

Since Michael Titelbaum's (2013) book is titled *Quitting Certainties*, it may seem that it would address certainty-reducing reasoning. However, Titelbaum's theory has nothing to say about the rationality of such reasoning. The title of Titelbaum's book is meant to signal that his theory does not imply that quitting certainties is irrational. Titelbaum tells us that his theory was primarily developed to address the limitations of simple conditionalization in addressing examples involving memory loss (or the threat of memory loss) and examples involving context-sensitive claims (2013, p. 4); though he also uses it as a framework for discussing examples involving what he calls the "expanding doxastic space"(88) and, as I discuss below, for addressing logical omniscience (106–110).

Why doesn't Titelbaum's theory enable us to evaluate the rationality of empirical certainty-reducing reasoning? The reason is that, for Titelbaum, in specifying an epistemic situation, the certainties are part of the extra-systematic constraints taken as *given* (40; 125 fn. 14). His theory "determines what requirements an agent's changing set of certainties places on her degrees of belief in claims of which she is less than certain"(40). Since Titelbaum's theory only places constraints on claims of which the agent is less than certain, and it does so, in part, on the basis of the claims of which she is certain, it has nothing to say about the rationality or irrationality of her certainties or the rationality or irrationality of certainty-reducing reasoning.

Titelbaum does allow that specifying the certainties in a particular case is a pragmatic decision, so we should consider the possibility that even though in the example above Crick and Watson change their attitude toward the evidence, even at the later time they are *pretty sure* that each item of evidence is true; so perhaps for practical purposes they can be modeled as certain of each item of evidence both before and after their discovery of the mistake made by their Cavendish colleagues.

I suggest this possibility only to be able to show why Titelbaum would surely reject it. On Titelbaum's account certainties are closed under deductive implication (41). Thus, if Titelbaum modeled the later time as a time at which Crick and Watson were certain of each item of evidence, his account would imply that Crick and Watson would have to be certain of the conjunction of all the evidence. But as the quotation from Crick makes clear, after the discovery of their colleagues' mistake, they were very confident that the conjunction of all of their evidence was false (or, at least, misleading). And their willingness to discard discrepant evidence could never be explained by a model that attributed to them certainty about the conjunction of all their evidence.

The Crick–Watson example does not raise problems for all Bayesian theories. Levi (1991) has a non-global Bayesian theory that can handle it, at least in theory. Levi allows for a confidence assignment of 1.0 to propositions whose negations are not serious possibilities (24). Levi's account of the hypothetical Crick and Watson case would be that their reasoning led them from originally not regarding mistaken evidence as a serious possibility to regarding the possibility that some item of evidence might be mistaken as a serious possibility (117–118). This would explain the change in their confidence assignments to the evidence. This reasoning would not be any form of conditionalization, but Levi allows for other forms of empirical reasoning; he is a non-global Bayesian. Shortly, I consider another example of a related kind of reasoning that could not be explained by Levi's account.

Cases like this one led Lewis (1980) and Jeffrey (1992) to a more radical conclusion: That a rational agent should never assign an extreme degree of confidence (0 or 1.0) to any empirical proposition. Of course, this immediately implies that no simple Bayesian view could be true, because simple conditionalization requires that evidence be certain. To solve this problem, Jeffrey introduced a principle of conditionalization that applies to uncertain evidence, *Jeffrey-conditionalization*.[1] Jeffrey-conditionalization cannot

[1] Jeffrey-conditionalization is the rule: Rational changes in degrees of confidence from prob_1 (at time t_1) to prob_2 (at a later time t_2) require that, where one's total change in evidence between t_1 and t_2 is given

explain certainty-reducing reasoning, but if agents don't ever regard their evidence as certain, they won't need to engage in certainty-reducing reasoning with respect to their evidence. In contrast, as I explain shortly, my non-Bayesian, non-probabilist principles can easily allow for certainty-reducing reasoning in cases such as this one and can explain what makes it rational (when it is).

2. Fallibility about deductive consequences. *Ian wonders: What is the trillionth digit in the decimal expansion of* π*? Ian knows only the first five digits in the decimal expansion of* π *(3.1416) but he has read authoritative reports that the decimal expansion of* π *is normal for at least the first 1.2 trillion digits—that is to say, for the first 1.2 trillion digits of the decimal expansion of* π*, each of the ten digits occurs with approximately the same relative frequency (.1). Since he has no other relevant knowledge about the trillionth digit in the decimal expansion of* π*, Ian assigns uniform confidence of .1 to each of the ten possible digits. Ten of Ian's friends each have a favorite digit: Ina (1), Dewey (2), Trey (3), Forest (4), Chico (5), Cecily (6), Stephen (7), Ada (8), Nina (9), and Sarah (0). Each of them has a strong hunch that the trillionth digit in the decimal expansion of* π *is their favorite digit. Each of them assigns confidence of 1.0 to their favorite digit and confidence of zero to each of the nine other possible values. Which of these eleven agents, if any, has a rational assignment of confidence to the ten alternatives?*

A strict classical probabilist holds that rationality requires that rational degrees of confidence satisfy the probability axioms and that they assign degree of confidence of one to all instances of all truths of classical logic. This implies a form of probabilistic closure:

> (*Classical probabilistic closure*) Whenever P logically implies Q (by classical logic), an agent's degree of confidence in Q must be at least as high as her degree of confidence in P.

As it happens, the trillionth digit in the decimal expansion of π is 2. Now consider:

(T) The trillionth digit in the decimal expansion of π is 2.

Since the conjunction of the axioms of mathematics (&AM), which, for simplicity, I assume include the definition of π, deductively imply T, the proposition [&AM $\supset$ T] is an instance of a truth of classical logic, so a strict classical probabilist would require Ian to assign confidence of 1.0 to that proposition; and, by probabilistic closure, Ian would be rationally required to assign at least as much confidence to T as he assigns to &AM, which, for simplicity, I assume to be 1.0. In fact, there are algorithms for determining whether T is true, but no human being could ever apply any of them in one lifetime. The only way we have of knowing whether T is true is to trust the results of computer applications of the algorithm.

Since we suppose that Ian is certain of &AM, on the strict, classical probabilist view he is required to be certain of T; so his confidence assignment to T is not rational. But, interestingly enough, Dewey's confidence assignment to T, which is based on

Footnote 1 continued
by a statement of the form $\mathrm{prob}_2(\mathrm{E}) = \mathrm{x}$: For any proposition P, $\mathrm{prob}_2(\mathrm{P}) = [\mathrm{prob}_1(\mathrm{P/E}) \times \mathrm{prob}_2(\mathrm{E})] + [\mathrm{prob}_1(\mathrm{P}/-\mathrm{E}) \times \mathrm{prob}_2(-\mathrm{E})]$ (cf. (e.g., Howson and Urbach 1989, p. 286).

a hunch, is rational on a strict, classical probabilist account. I think it is clear that on any acceptable human conception of rationality, Ian's confidence assignment to T is rational and Dewey's is highly irrational. In addition, I think it is clear that any acceptable human conception of rationality would require that Dewey, who clearly recognizes that his confidence assignment to T is based on nothing more than a hunch, reduce his confidence in T from 1.0 to .1.

Unlike the previous example, this example raises problems for Levi's (1991) account. First, Dewey's certainty-reducing reasoning with respect to T cannot be rational on Levi's account. Recall that Levi allows for agents to change their confidence assignments from extreme to non-extreme values when an alternative goes from not being to being a serious possibility. However, since Levi is a strict classical probabilist, that T could be false could never be a serious possibility for Levi. It would not be rational, on Levi's account, for Ian to assign confidence of .1 to T; nor would it be rational, on Levi's account, for *any* of Ian's friends to reduce their certainty in their guesses, because to do so would violate their confirmational commitments, since they are *all* committed to assigning confidence of one to *some* value (1991, pp. 52–53).

Garber (1983) and Jeffrey (1983) have different but similar ways of addressing related cases. They propose that we employ a formalism that, in addition to representing the mathematical axioms and propositions stating the ten alternative values for the trillionth digit in the decimal expansion of π, includes logical atoms of the form [P $\vdash$ Q]. The idea is that if we regard formulas that fit the schema [P $\vdash$ Q] as uninterpreted formulas satisfying certain judiciously chosen formal constraints, then we can use the formalism to model an agent who is uncertain of valid deductive relations and who comes to assign high confidence to [P $\vdash$ Q] on the basis of empirical evidence (e.g., the results of a computer computation).

Jeffrey and Garber implement this idea in different ways. Jeffrey implements it into a strict classical probabilist framework, which enables him to model some discoveries of logical relations, but not in a way that will help with this case, because to explain the rationality of Ian's confidence assignment to T, it is necessary to explain not only how Ian could assign confidence of less than 1.0 to [&AM $\vdash$ T], which Jeffrey can very well explain, but also to explain how it could be rational for Ian to assign less confidence to T than to &AM, which no strict classical probabilist account such as Jeffrey's could explain, because it would violate probabilistic closure.[2]

Garber's (1983) proposal is more promising, because he abandons strict classical probabilism for a version of non-strict classical probabilism and only requires that all tautologies (i.e., truth functionally valid formulas) be assigned confidence of 1.0 (Garber 1983), p. 113). Since [&AM $\supset$ T] is not truth functionally valid, Garber's account does not imply that Ian's probability assignment to T must be at least as high

[2] The technical explanation of this result is that, for Jeffrey, rational degrees of confidence are sets of confidence assignments, each element of which satisfies the standard probability axioms (1983, pp. 143, 145–146). That Jeffrey is committed to probabilistic closure is undeniable, because he uses it in one of his proofs (1983, p. 152). The problem that Jeffrey's proposal was meant to address, the problem of old evidence (e.g., Glymour 1980), is different from the current example, because the problem of old evidence does not involve a violation of probabilistic closure. (At all times in the process by which a theory comes to acquire confirmation by old evidence, the old evidence is more probable than the theory.)

as his probability assignment to &AM. But it clearly generates problems of a similar kind.

Consider Chris, who wonders whether an extremely complex truth functional proposition P is truth functionally valid. As in Ian's case, there are algorithms for determining whether P is truth functionally valid, but, depending on how complex P is, there may be no way for a human being to use any of them to get an answer in a single human lifetime.

If Chris does not know the results of applying the algorithm to P, I believe that the most rational thing for Chris to do would be to assign a non-extreme degree of confidence to P. However, suppose that P is truth functionally valid and Dan just has a hunch that it is, so Dan assigns confidence of 1.0 to P. In Garber's system, Dan's assignment of confidence to P would be rational and Chris's would not be. We need a principle that reverses this verdict. There is also another problem for Garber's proposal that I take up shortly.

Seidenfeld et al. (2012) argue against Garber's proposal and in favor of one that ranks degrees of incoherence of probability assignments, where degrees of incoherence are based on a measure of how far the relevant degrees of confidence depart from the requirements of strict classical probabilism. For this reason, I classify their proposal as a strict classical probabilist one. Whatever other merits their proposal may have, it does not help with these cases, because it implies that Dewey's confidence assignment of 1.0 to T is more rational than Ian's assignment of .1; and it implies that Dan's confidence assignment of 1.0 to P is more rational than Chris's non-extreme assignment of confidence to P. Again, we need a principle that reverses both verdicts. Shortly, I introduce principles that do so.

Titelbaum develops a modeling framework that can model non-probabilist degrees of confidence (2013, pp. 106–110). The confidence assignments of Ian and Dewey can be modeled in Titelbaum's framework. But the framework is so permissive that it could hardly explain the irrationality of *any* changes in degrees of confidence; so, after introducing the general framework, Titelbaum immediately returns to a strict classical probabilist version of it (2013, pp. 109–110). We need to find a more useful non-probabilist framework.

3(a). Fallibility about the laws of logic generally. Sharon reads Quine's (1961) *"Two Dogmas of Empiricism," in which Quine presents a historical argument based on the history of examples of quitting certainties in every area of inquiry and she decides, on the basis of that historical evidence, that she should not be certain of the laws of logic. Then she reads the argument of* Field (1996) *that the laws of classical logic are special in that our evidential system makes them immune to empirical disconfirmation. She finds especially compelling Field's argument that any evidence-based reasoning for a change in logic would have to have logic built into it (369). She notices that Quine's argument itself employs classical logic. How could it be rational to use classically logical arguments based on empirical evidence to conclude that one should be less than fully certain of the truths of classical logic?*

I discuss the issue abstractly in this example and more particularly in the next. Strict classical probabilist views require that all instances of truths of classical logic be assigned confidence of 1.0 and even Garber's (1983) non-strict classical probabilism requires that all instances of the law of non-contradiction and the law of excluded middle be assigned confidence of 1.0. They could not explain how empirical, historical information could make it rational for Sharon to reduce her confidence assignment to instances of the law of excluded middle or the law of non-contradiction to less than 1.0.[3] Is Field's (1996) argument an adequate defense of probabilism? It is not necessary to address the details of Field's argument in order to address the most important issue raised by it: If Sharon is familiar with the history of philosophy, could any philosophical argument make it rational for her to assign confidence of 1.0 to any or all of the laws of classical logic? I return to that question below.

3(b). Fallibility about specific laws of logic. Maya becomes interested in the theory of truth. She reads that there are three main categories of such theories: (1) classical theories (e.g., Tarski 1944*) that employ the laws of classical logic; (2) paracomplete theories (e.g.,* Kripke 1975; Field 2008*) that allow for exceptions to the law of excluded middle; and (3) paraconsistent theories (e.g.,* Priest 2002*) that allow for exceptions to the law of non-contradiction. Maya recognizes that Kripke, Field, and Priest are all prominent philosophers. Before reading any of their theories, what degree of confidence should Maya assign to the proposition that in the correct theory of truth there are exceptions to the law of excluded middle or to the law of non-contradiction?*

Here, we have a specific challenge both to strict classical probabilism and to Garber's non-strict classical probabilism. Also here, if not before, we have reason to wonder whether those 'idealized' agents of Bayesian and non-Bayesian probabilist theories really are rationally ideal. A strict classical probabilist agent or a non-strict classical probabilist agent of the Garber variety would assign degree of confidence of 1.0 to all the instances of the law of excluded middle and the law of non-contradiction, so she would assign degree of confidence of zero to the paracomplete theories of Kripke and Field and to the paraconsistent theory of Priest. Only Tarksi's theory would receive a positive assignment of confidence. I take it that it requires no familiarity with the theory of truth to recognize that if such prominent philosophers as Kripke, Field, and Priest (among others) have seriously proposed nonclassical theories, it could not be rational to assign degree of confidence of zero to their theories. Purely on the basis of empirical information concerning the philosophical reputations of Kripke, Field, and Priest, Maya should assign some positive degree of confidence to their theories.

In addition, anyone who assigned degree of confidence of zero to their theories would be in the position of holding that evidence of the prominence of Kripke, Field, and Priest in philosophy would be irrelevant to whether it would be rational to read what they say in defense of their theories to find out if there were any good reasons for adopting them, because there *could be* no good reason for adopting them. So it is a problem for all Bayesian and non-Bayesian classical probabilist theories that none of them could explain how it could be rational for Maya to assign any non-zero degree of confidence to the theories of Kripke, Field, or Priest; or how it could be rational for Maya to use empirical evidence to support empirical confidence-reducing reasoning from a prior confidence assignment of 1.0 to the laws of classical logic.

[3] I emphasize that the certainty-reducing reasoning in this example (and in all the others that I discuss) is empirical reasoning based solely on empirical evidence, to forestall the potential objection that Bayesianism does not apply to a priori reasoning.

4. Fallibility about infallibility (with an application to epistemology). On Monday, Immanuel publishes his epistemological treatise T_1 in which he asserts that T_1 is itself validated by his infallible a priori faculty (which gives him infallible insight into the non-empirical, metaphysically necessary truths of epistemology), and he lists 1000 epistemological claims validated by his a priori faculty.

On Tuesday, Immanuel publishes T_2, in which he acknowledges that some of the statements in T_1 were false, but he asserts that he has corrected the flaw in Monday's a priori reasoning and now, relying only on his infallible ***special*** *a priori faculty, he asserts T_2, which includes only 900 of the 1000 claims included in T_1.*

On Wednesday, Immanuel publishes T_3, in which he acknowledges that some of the statements in T_2 were false, but he asserts that he has corrected the flaw in Tuesday's special a priori reasoning and now, relying only on his infallible ***very special*** *a priori faculty, he asserts T_3, which includes only 800 of the 900 claims included in T_2.*

On Thursday, Immanuel publishes T_4, in which he acknowledges that some of the statements in T_3 were false, but he asserts that he has corrected the flaw in Wednesday's very special a priori reasoning and now, relying only on his infallible ***very, very special*** *a priori faculty, he asserts T_4, which includes only 700 of the 800 claims included in T_3:*

At least by Thursday (and almost certainly before), it is irrational for Immanuel to assign confidence of 1.0 to the propositions of his theory T_4, even if, mirabile dictu, his very, very special a priori faculty is infallible. Why? Let us say that an epistemology that asserts or implies that rationality requires assigning confidence of 1.0 to any of its substantive (i.e., non-trivial) claims is immodest. Immanuel's epistemology is immodest. Given the history of mistakes in epistemology, no immodest epistemology could be rational. Why not?

The example of Immanuel is intended to provide a microcosm of the history of Western epistemology, which has included far too many epistemologists whose views implied that their epistemologies were rationally certain and whose epistemologies were *self-insulating*—that is, whose views implied that all those who disagreed with them were irrational. In the history of epistemology, we are well past Thursday, so we have plenty of historical evidence that rationally undermines such claims. Shortly, I explain the empirical certainty-reducing reasoning that does so.

This example raises a serious issue for Bayesians and other probabilists, because even if, as I assume, they would agree that by Thursday Immanuel's confidence assignments of 1.0 to the propositions of his epistemology are irrational, they themselves are committed to asserting that rationality requires confidence of 1.0 in at least some of their own non-trivial epistemological claims. Probabilism implies that rationality requires assigning confidence of 1.0 to at least some instances of logical truths. Let L_1 be one of those instances. Probabilism implies that rationality requires assigning confidence of 1.0 to the epistemological proposition that rationality requires assigning confidence of 1.0 to L_1.[4] Thus, probabilist epistemology is immodest. But the same kind of reasoning that undermines Immanuel's certainty in his epistemology on Thursday also undermines certainty in the epistemological claims of probabilism.

[4] See the Appendix for a proof.

I believe that if there is anything we have learned from the history of epistemology, it is that it could no longer be rational to assign confidence of 1.0 to any substantive epistemological claim. And for substantive epistemological *principles*, the situation is even worse. Given the history of mistakes and disagreement in epistemology, a rational confidence assignment to a substantive epistemological principle would almost surely be less than .5. In the history of epistemology, I know of no epistemological principle that could explain why it was rational for us to assign it confidence of less than .5, until now.

Shortly, I introduce a principle that can explain not only why, on Thursday, it is irrational for Immanuel to assign confidence of 1.0 to his epistemology, but also why, given our evidence, any immodest epistemology is irrational. My argument is based on empirical evidence from the history of epistemology. So it is an empirical argument. If there is any hope for probabilist epistemology, there must be something wrong with my argument. Titelbaum (2015) thinks there is. I consider his argument after I have explained one of my principles and applied it to the examples.

5. Fallibility about everything. Once we use empirical evidence to conclude that we are fallible about all the items in examples (1)–(4) above, it is a short step to concluding that we are fallible about everything, including that very claim (Christensen 2007).

Clearly, Bayesian and non-Bayesian probabilist accounts could never explain this result. In the history of epistemology, before now, has there ever been an epistemological principle that could explain how it could be rational not to assign confidence of 1.0 (or 0) to anything, not even to it?

3 The special metacognitive expected relative frequency (MERF) principle

In this section, I present a principle that, when suitably generalized, explains all of the examples above. The principle is a principle of epistemic rationality, not pragmatic or some other kind of rationality. It is a principle of higher-order reasoning, which means that it is not involved in ground-level determinations of rational beliefs and degrees of confidence. It typically comes into play when something triggers a question about the reliability of those ground-level determinations.[5] To call attention to the difference between ground-level and meta-level processing, I introduce a harmless fiction by distinguishing between the self who does the ground-level cognitive processing—the *ground-level self (GLS)*—and the self who does the metacognitive processing—the *metacognitive self (MCS)*.

One important principle governing the MCS's reasoning is a metacognitive expected relative frequency (MERF) principle.[6] Although a precise statement of the principle involves some complexity, the main idea of it is quite intuitive. When a question arises about the reliability of ground-level confidence assignments, the metacognitive self

[5] Though, as I explain the Appendix, the reliability of metacognitive processes can be evaluated in the same way.

[6] There is also a ground-level expected relative frequency (GERF) principle, but I ignore it here. It is clearly beyond the scope of this paper to consider the principles that govern ground-level cognitive processing.

(MCS) steps back and evaluates the relevant ground-level processes by, to use a term from Christensen (2011), *bracketing* the ground-level judgments. The MCS considers only certain kinds of information in its determination of the reliability of the processes responsible for the ground-level confidence assignments; it does not simply repeat the ground-level processing. The MCS then makes an estimate of the relative frequency of truth of the relevant ground-level processes and that estimate is the basis for a recalibration of the relevant degrees of confidence, if one is needed.

Before stating the MERF principle and applying it to the preceding examples, let's consider a simple example that can be used to explain the main ideas:

6. The fallible calculator. At restaurants, David generally tips 18 % of the bill. Today David is having dinner by himself at a restaurant. After dinner he calculates the tip to be $4.59 and he is practically certain of the result. He assigns confidence of .99 to the result.

For simplicity, I suppose that David's MCS classifies the causal processes involved in producing the result into a kind (e.g., manual calculation of simple products) and the MCS has enough information on David's past manual calculations of simple products to estimate that they have been correct 95 % of the time. David's MCS then estimates the expected relative frequency of truth in calculations of that kind (a mathematical expectation of the relative frequency of truth). In a simple case like this one, the MCS can determine an expected relative frequency of truth for the relevant kind of process by simply projecting the past relative frequency of truth into the future. For David's confidence assignment of .99 to the mathematical proposition stating the result of his calculation to be in metacognitive equilibrium, the MERF principle requires that it be equal (or approximately equal) to the MCS-determined expected relative frequency of truth of propositions assigned confidence of .99 by the relevant cognitive process. In this case, that expected relative frequency is .95; since $.95 \not\approx .99$, with one qualification that I discuss in the next paragraph, David's confidence assignment is in metacognitive disequilibrium. If so, then metacognitive rationality requires a change in the ground level assignment or the relevant metacognitively determined expected relative frequency of truth (or both) to make them equal (or approximately equal) in order to achieve equilibrium. The MERF principle does not by itself determine what changes should be made to achieve equilibrium. In all the cases I discuss here, I will assume that rationality requires that the ground-level determination of confidence should be modified to bring it in to equilibrium with the metacognitive determination. So in this case, with one qualification that I discuss in the next paragraph, rationality requires David's MCS to reduce his confidence in the results of his calculation from .99 to .95.

The qualification is this: The results of the metacognitive analysis depend on the way that the relevant causal processes are categorized. More precise reliability-relevant specifications of the causal processes responsible for an agent's confidence assignment to a proposition P can change the expected relative frequency of truth that is relevant to whether the agent's confidence assignment to P is in equilibrium. Suppose that it was true that David did a manual calculation, but that he also checked that calculation on a calculator. And suppose that David's MCS estimates the relative frequency of truth of calculations that are done manually and checked on a calculator to be .99. Then David's original confidence assignment of .99 to the result of his calculation would be in metacognitive equilibrium, unless there were some further reliability-relevant specification of the relevant kind of causal process that led David's MCS to make yet a different determination of the expected frequency of truth among propositions assigned confidence of .99 by cognitive processes of that further specified kind. This process of adding further reliability-relevant specifications of the kind of causal process involved is an illustration of what I call the *Narrower Reference Class Exception* that is part of MERF. In theory, there is no limit to the number of exceptions and exceptions to exceptions countenanced by the Narrower Reference Class Exception. In practice, the exceptions bottom out when the MCS judges that there are no further relevant sub-categories.[7]

What is the MERF principle? I begin with a special, limited version of the MERF principle, which is stated in terms of the metacognitive *beliefs* of the MCS. The final version of the MERF principle, generalized to replace metacognitive beliefs with metacognitive degrees of confidence, appears in the Appendix. To state the principle more perspicuously, I introduce an abbreviation:

> $ERF(T_r/conf = c, CP_1)$ = the expected relative frequency of truths among propositions assigned confidence of c on the basis of a causal process of kind CP_1.

Here is the special version of the principle:

> *(Special metacognitive expected relative frequency (MERF) Principle) An agent's confidence assignment of c to a proposition P is in disequilibrium if: The MCS judges there to be a reliability-relevant categorization CP_1 of the causal processes responsible for the agent's confidence assignment of c to P, such that:*
> $ERF(T_r/conf = c, CP_1) = d$
> *and it is not the case that* $d \approx c$;
> *UNLESS [Narrower Reference Class Exception] the MCS believes that there is a reliability-relevant categorization CP_2 of the causal processes responsible for the agent's confidence assignment of c to P*, $CP_2 \leq CP_1$, *such that:*
> $ERF(T_r/conf = c, CP_2) \approx c$.[8]

[7] Thus, there is no analog of reliabilism's generality problem here (e.g., Feldman 1985), because the MCS's determination of the relevant expected relative frequencies of truth are always relative to the information available to the MCS, which is assumed to be finite.

[8] The special MERF Principle is a refinement of an earlier proposal (Talbott 1990, Text 110–112). I should mention here that, to avoid unnecessary complications, the version of the MERF principle stated in the text assumes that all of the MCS's judgments of relative frequency of truth are *well-behaved*, in the following sense: When an MCS evaluates the reliability of the agent's assignment of confidence of c to proposition P, if the MCS believes there to be reliability-relevant categorizations of cognitive processes CP_1 and CP_2 that both include the cognitive processes responsible for the agent's assignment of confidence of c to P, neither of which is a proper sub-class of the other; and if the MCS believes that $\mathrm{ERF}(\mathrm{T_r/conf{=}c}, \mathrm{CP}_1)$=d and that $\mathrm{ERF}(\mathrm{T_r/conf{=}c}, \mathrm{CP}_2)$=e and it is not the case that d $\approx$ e, then the MCS also has a belief about the value of the $\mathrm{ERF}(\mathrm{T_r/conf{=}c}, \mathrm{CP}_1\ \&\ \mathrm{CP}_2)$. This requirement can be illustrated by a simple example. Suppose that an agent assigns confidence of c to a proposition P (the result of an arithmetical calculation) on the basis of having performed two different algorithms (CP_1 and CP_2), which both agreed on the solution. If the agent's MCS has an opinion about the expected relative frequency of truth when each of these algorithms is applied individually—that is, an opinion about $\mathrm{ERF}(\mathrm{T_r/conf{=}c}, \mathrm{CP}_1)$ and $\mathrm{ERF}(\mathrm{T_r/conf{=}c}, \mathrm{CP}_2)$—then the MCS also has an opinion about the expected relative frequency of truth when the combination of algorithms ($\mathrm{CP}_1\&\mathrm{CP}_2$) is applied (i.e., $\mathrm{ERF}(\mathrm{T_r/conf} = \mathrm{c}, \mathrm{CP}_1\&\mathrm{CP}_2)$). I discuss this assumption of well-behavedness more fully below.

It is important to recognize that the special MERF principle is a principle that *governs* rational metacognition, not a principle that the MCS (or anyone else) explicitly applies. The MCS makes the relevant categorizations (usually unconsciously) and then the special MERF principle defines whether they are in equilibrium. If they are in disequilibrium, rationality requires a recalibration of the agent's degrees of confidence to bring them into equilibrium. This process is typically unconscious and automatic.

In evaluating the reliability of an agent's confidence assignment to a proposition P, the agent's MCS brackets the agent's confidence assignment to P, and focuses instead on information about the causal processes responsible for the confidence assignment to P. Only some categorizations of causal processes (CP) are relevant to these meta-level determinations of reliability. I refer to them as the *reliability-relevant* categorizations.[9]

Notice that the special MERF principle does not require that agents make *any* determinations of the reliability of their confidence assignments. The confidence assignments of young children are in rational equilibrium by default, because young children cannot evaluate their own reliability. The MERF principle only comes into play when an agent does make such determinations and, to put it intuitively, her ground-level confidence assignments to some propositions do not match the MCS's estimate of their expected reliability.

I have already explained how the MERF principle applies to the example of the fallible calculator. The same kind of reasoning is implicated when, after remembering that I have recently forgotten some of my appointments, I place less confidence in the proposition that my afternoon is open when it is based on my failure to remember any scheduled events; or when, after reading research on the unreliability of eyewitness identifications by experimental subjects, I reduce my confidence in my own eyewitness identifications; or when, after reading research on the unreliability of rankings of job candidates based on personal interviews, I reduce my confidence in my rankings of job candidates based on personal interviews. There are almost limitless examples of this kind of metacognitive recalibration. Many of them can be explained within a Bayesian framework, but only a non-Bayesian, non-probabilist framework can explain all of them.

I should also say that, although the MERF principle explains the certainty-reducing reasoning in the examples I discuss here, it does not explain all certainty-reducing reasoning. Consider a case in which Allen starts out being certain of p because he is certain that −p is impossible. If a process of ground-level reasoning leads him to understand how −p could be possible, this ground-level process will also reduce his certainty in p. In this case, Allen's certainty-reducing reasoning is ground-level, not metacognitive, so the MERF principle does not even apply to it. This is just a reminder that my MERF principle is not the only principle of epistemic rationality. It is not even the only principle of metacognitive rationality.

[9] For example, a reliability-relevant categorization of David's belief about the amount to be tipped is that it is the result of a manual calculation of a simple arithmetical product. A categorization that is not relevant to a determination of its reliability would be that it is a *correct manual calculation*, even though David's GCS may believe that it is a correct manual calculation. For more on the distinction between those kinds that are reliability-relevant and those that are not, see Talbott (1990) (Preface 21–30 and Text 81–96).

Let's now apply the special MERF principle to the examples discussed above.

3.1 Fallibility about evidence

Consider an elaboration of the Crick-Watson example. Suppose that Crick and Watson obtained X-ray diffraction patterns of the structure of DNA and used those patterns to calculate the structure of various elements of the DNA molecules. Suppose they made many calculations and regarded each of them as practically certain (confidence of 1.0)—that is, they did not consider the possibility of a mistaken calculation to be a serious possibility.

However, suppose that they later found out that their colleagues failed to discover the structure of one form of a polypeptide chain, because they had regarded their calculations as practically certain, when in fact 1 % of them were mistaken. And then suppose that they undertook further investigations and found out that other colleagues had made similar mistakes in the past, with approximately the same error rate. Finally, suppose that they have good reason to believe that their calculations are neither more nor less reliable than those of their colleagues. In order to work out the structure of DNA, it has been necessary for them to make 1000 such calculations. Each calculation seems to them to be practically certain and before obtaining the information about the fallibility of their colleagues' calculations, they were practically certain (confidence of 1.0) of the results of each calculation, because they did not regard an erroneous calculation as a serious possibility. In light of the evidence on reliability, what should their MCS's attitude be toward the results of the 1000 calculations?

Let CP_1 be a reliability-relevant classification of the causal processes involved in making such calculations. Let P be a proposition stating the result of one such calculation. On reasonable assumptions, the MCS should determine the expected relative frequency of truth of calculations of this kind, which they regard as practically certain (confidence of 1.0) to be .99, that is:

$$\mathrm{ERF}(\mathrm{T_r}/\mathrm{conf} = 1.0, \mathrm{CP}_1) = .99.$$

If there is no further reliability-relevant specification of the kind of causal processes involved, as there would be, for example, if some of the calculations were quite complex and some were simple calculations that they could do in their head, their confidence assignments to each of the results of the 1000 calculations are in metacognitive disequilibrium until by an automatic process, their confidence assignment to each of the 1000 results is reduced to .99, and equilibrium is restored.

This is a relatively simple example, but it provides a model of empirical certainty-reducing reasoning that can be applied quite generally. There are no limits on its domain of applicability.

3.2 Fallibility about deductive consequences

In the example of Ian's fallibility about deductive consequences, there were two things that needed explaining: (1) the rationality of Ian's assignment of confidence of .1 to

T; (2) the rationality of Dewey's certainty-reducing reasoning that would lead him to reduce his confidence in T from 1.0 to .01. Since an explanation of (2) is also an explanation of (1), I discuss (2).

When Dewey's MCS evaluates the reliability of Dewey's hunches, I assume his MCS has lots of evidence of mistaken hunches. Dewey may feel sure of this hunch, but his MCS can see that each of the other nine agents are just as sure of their hunches. On almost any reasonable assumptions, given that the decimal expansion of π is normal for the first 1.2 trillion places, if Dewey's MCS is rational, it will evaluate the expected relative frequency of truth his hunches (H) of certainty to be the same as for the other agents'—that is, .1:

$$\mathrm{ERF}(\mathrm{T_r}/\mathrm{conf} = 1.0, \mathrm{H}) = .1$$

In the absence of any further specification of the causal processes involved that would change the expected relative frequency of truth relevant to this case, Dewey's confidence assignment to T would be in metacognitive disequilibrium. To restore equilibrium, there would be an automatic adjustment in his confidence assignment to T to make it equal to .1.

3.3 Fallibility about the laws of logic generally

At first, Sharon is persuaded by Quine's historical argument that she should revise her confidence assignments to the instances of the laws of classical logic to something less than 1.0. Then she reads Field's argument and wonders how it could be rational to assign confidence less than 1.0 to the laws of classical logic on the basis of an argument using the laws of classical logic (Field 1996, p. 369). What degrees of confidence should she assign to the instances of the laws of classical logic?

If Sharon is familiar with the history of philosophy, no matter how compelling Field's argument appears to her to be, her MCS will not be able to rationally assign confidence of 1.0 to the laws of classical logic on the basis of it. The hypothetical example of Immanuel above is directly relevant here. On each day of the week after Monday, Immanuel acknowledges that his previous argument for a special class of infallible beliefs failed and he defines a narrower subclass and argues that the members of the subclass are really special. In that example, all of the subclasses were defined by one person. In the actual history of philosophy, they have been defined by successive philosophers. But the rational considerations are the same.

Field is making his argument well past Thursday. Just as it is theoretically possible that on Thursday Immanuel has actually hit upon a very, very special faculty of a priori intuition that really is infallible, it is theoretically possible that Field has hit upon a correct argument for the conclusion that our evidential system is one in which no empirical evidence could disconfirm an instance of a law of classical logic; but it would not be rational for anyone familiar with the history of philosophy (which includes Field) to be certain of either claim. So Sharon (and Field) should assign confidence of less than 1.0 to Field's conclusion that she should be certain of the laws of classical logic, and, as a result, there is nothing to block her from assigning confidence of less than 1.0 to the laws themselves. Though no probabilist account can explain this, the special MERF principle easily explains it. In a case like this, Sharon's MCS will estimate the expected frequency of truth among her confidence assignments of 1.0 to instances of the laws of classical logic to be unreliable by a miniscule amount (ϵ). This will introduce a disequilibrium into her confidence assignment that will be resolved by reducing her confidence in each of the instances of the laws of classical logic to $(1 - \epsilon)$.

The only remaining question is to explain how it could be rational to assign confidence of less than 1.0 to the laws of classical logic on the basis of an argument employing the laws of classical logic. A full answer would require an alternative account of reasoning that did not require that the laws of classical logic have probability 1.0. In this paper, I only try to articulate one part of such an account. But even without a full account, the analogy to Euclidean geometry is helpful. For Euclidean geometry to be useful, it did not have to be literally true, only—to employ an idea from Millgram (2009)—*true enough* to give true or approximately true or true enough results in ordinary applications. Similarly, for classical logic to be useful in arguments, it is not necessary that it be literally true, only true enough to give true or approximately true or true enough results in ordinary applications.

But isn't it true that the two versions of the MERF principle are part of an epistemological theory that has logical laws? Won't the theory have to assign confidence of 1.0 to those logical laws? It is true that my epistemological theory has logical laws, but, unlike probabilism, my epistemological theory does not impose its logical laws on the epistemically rational agents that it is a theory of.[10] The special and general versions of the MERF principle are part of an overall epistemological theory that includes logical laws, but the theory can explain how it could be rational for an agent to endorse a different logic.

3.4 Fallibility about specific laws of logic

It would take us too far afield to try to rationally evaluate the various nonclassical proposals for a theory of truth (e.g., Kripke 1975; Priest 2002; Field 2008). Fortunately, we do not have to do so and neither does Maya for the special MERF principle to come into play. We only need the empirical evidence available to Maya to trigger the MERF certainty-reducing reasoning.

Let L be a typical liar sentence (e.g. "This statement is not true"). All Maya needs to know is that prominent philosophers such as Kripke (1975) and Field (2008) have advocated the view that the instance of the law of excluded middle [L is true v -L is true] is not true, for it to be rational for Maya to assign it confidence of less than 1.0; and all she needs to know is that prominent philosophers such as Priest (2002) have advocated the view that the following explicit contradiction is true: [L is true & -L is not true], for it to be rational for Maya to assign it confidence of greater than zero. The special MERF principle easily explains this result, because Maya's MCS

[10] An interesting question, which I can only mention here, is this: Is the logic of my epistemological theory classical? Because the concept of truth plays a crucial role in my epistemological theory and because I am persuaded that an adequate theory of truth requires a nonclassical logic, I believe that my epistemological theory requires a nonclassical logic. But classical logic is true enough for the purposes of this paper.

would acknowledge that the relative frequency of truth of propositions that have been defended by prominent philosophers even though they were at one time intuitively unattractive is greater than zero.

Probabilism has clearly irrational implications in this example. Probabilism requires assigning confidence of zero to the positions of Kripke or Field or Priest. To do so, one would have to think there would be no reason to even read what they had to say in favor of their positions. Thus, a probabilist agent would be closed-minded in a way that seems far from an ideal of rationality.

3.5 Fallibility about infallibility (with an application to epistemology)

On the first day (Monday) Immanuel identifies an a priori faculty for apprehending truths of epistemology and identifies 1000 propositions as infallibly true on the basis of that faculty. On each successive day, Immanuel acknowledges making an error on the preceding day, but defines a new infallible a priori faculty for apprehending a narrower set of truths of epistemology. Purely empirical information about the history of his mistakes makes it irrational by Thursday, if not before, for Immanuel to regard any of his a priori faculties as infallible (no matter what they tell him). This introduces a metacognitive disequilibrium that can only be rationally resolved by reducing his confidence to all epistemological propositions antecedently regarded as certain on the basis of any of his a priori faculties.

The example applies directly to all Bayesian and non-Bayesian probabilist epistemologists, because Immanuel is a microcosm of the history of Western epistemology and probabilist epistemologists are making claims to rational certainty for their epistemology well past Thursday. For anyone who is familiar with the history of Western epistemology, I don't see how it can be rational to assign confidence of 1.0 to any substantive epistemological claim. Any rational epistemology must be modest. The MERF principle easily explains this result.

For substantive epistemological principles, the situation is even more dire. Having recognized that there may even be exceptions to the least controversial principles—that law of noncontradiction and the law of excluded middle—I do not believe that it could be rational to assign confidence above .5, or even close to .5, to any controversial epistemological principle. But this implies that it is not rational for me to assign confidence of .5 to my own MERF principle, because the MERF principle will be a very controversial principle. Could the MERF principle explain that?

The answer is "Yes," but to understand how it could be "Yes," it is necessary to remember that no one, not the agent, not her GLS, not her MCS, *applies* the MERF principle in the reasoning governed by the principle. The GLS determines ground level confidence assignments; the MCS determines metacognitive confidence assignments; the MERF principle defines when they are in metacognitive disequilibrium; and when they are, the agent's cognitive system operates automatically to restore equilibrium. Thus, the MERF principle can explain why David's confidence assignments of .99 to the results of manual arithmetical calculations that he has checked on a calculator is in metacognitive equilibrium, even if David has no idea that there is any such principle.

In the determination of how much confidence I should place in the MERF principle, my MCS will review the history of epistemology, a history largely of failed efforts to provide exceptionless, substantive epistemological principles and will estimate the expected relative frequency of truth to be low. My MCS won't be able to make any kind of persuasive case that my epistemological reasoning is more reliable than theirs. I do think that a reasonable case can be made that in the history of epistemology there has been some improvement in the principles proposed, to which, I believe, the example of probabilist epistemology lends support. I think that there are many dimensions on which probabilist epistemology is a substantial improvement over earlier epistemologies. But the relative frequency of true, exceptionless principles in the history of epistemology is too low to make it rational for me to assign confidence anywhere near .5 to my own MERF principle. That is the MERF principle's own verdict on itself.

I will say that an epistemological principle that implies that, given our evidence, our degree of confidence in it should be less than .5 (and thus, that we ought not to believe it) is *extremely modest*. Given the history of mistakes and disagreement in epistemology, it would seem to be a condition of adequacy on any substantive epistemological principle that applies to itself that it be extremely modest. So far as I know, the two versions of the MERF principle are the first substantive epistemological principles to be extremely modest.

Echoing Laudan's (1981) argument about science, we might call the conclusion that we should not believe any substantive epistemological principle to be exceptionless the *pessimistic epistemological induction*. It is crucial to recognize that, to the extent that the reasoning involved in this argument is inductive, it is not just another instance of ordinary, ground-level, inductive reasoning, because it involves metacognitive *defeat* of the results of ordinary, ground-level reasoning. The kind of defeater involved is what Pollock called an *undercutting defeater* or, more precisely, a *reliability defeater* (1984, pp. 112–113).

Could rationality really require us to make the pessimistic epistemological induction? Perhaps I am making a mistake here. Perhaps there is a kind of epistemological proposition that is indefeasible and, thus, not subject to metacognitive defeat governed by the MERF principle. As it happens, Titelbaum (2015) has made an ingenious argument for just that claim. I take up Titelbaum's argument for his no-reliability defeater-claim shortly.

3.6 Fallibility about everything

Lewis (1980) and Jeffrey (1992) advised us not to be certain of any empirical matter of fact. Both Lewis and Jeffrey were probabilists, so they could not extend their advice to the instances of the laws of classical logic; nor could they extend it to their own probabilist epistemology. Is there any epistemological principle that can explain how it could be rational to be certain of nothing?[11] Not even the special MERF principle can explain this result, because it only applies to an MCS who is able to hold metacognitive *beliefs*. To solve this problem, I need a more general version of the MERF principle,

[11] Here I only insist that it *could* be rational to be certain of nothing. For more considerations that could support this conclusion, see Christensen (2007). To argue that it is rational to be certain of nothing, I would have to consider more examples—for example, the Cartesian examples: *I am thinking* or *I exist*.

one that operates on metacognitive confidence assignments to the relevant meta-level propositions.

The general MERF principle will make it possible (and potentially required) for me to assign non-extreme degrees of confidence (i.e., strictly between 0 and 1) to every proposition, including the proposition that I do have a source of infallible rational insight and including the proposition that I do not assign confidence of 1.0 to any proposition, because I should not be certain that rationality requires me to assign confidence between 0 and 1 to every proposition and, even if it does, I should not be certain that my confidence assignments satisfy that requirement. I leave the details to the Appendix.

4 Defenses of probabilism

In this paper I have made a case for a non-probabilist principle of metacognitive rationality based on examples that raise problems for Bayesianism and for classical and non-classical probabilism. But there are many defenses of probabilism that argue that violating the standard probability axioms must be irrational. In this section, I review the most influential of these arguments. In a sense, I have already replied to these arguments, because each of the examples discussed above is an example of how to rationally violate probabilist or Bayesian conditions on epistemic rationality. Here I review the arguments, not so much to show where they go wrong, as to show how to restate what they get right in a non-probabilist framework.

Dutch Book defenses of classical probabilism (e.g., de Finetti [1937]), are arguments that the failure to satisfy the standard probability axioms leaves one open to a Dutch Book of wagers—that is, a combination of wagers that the agent is sure to lose on the basis of logic alone.[12] Paris (2001) has shown how to extend these defenses to provide Dutch Book defenses of non-classical probabilism. There are many problems with treating degrees of confidence as betting odds, but here I set all those problems aside.

Consider again the example of Ian and the trillionth digit of π or the example of Chris who was in doubt about truth functional validity. For any probabilist account, there will be examples like these. No human conception of rationality would classify Ian or Chris's confidence assignments as irrational, even though they are susceptible to a Dutch Book on the basis of logic alone. The MERF principle explains why their confidence assignments are rational.

A different kind of argument for classical probabilism is due to Joyce (1998). Joyce believes that just as full belief has a purely epistemic goal of truth, we can formulate a purely epistemic goal for partial belief: gradational accuracy. Joyce actually focuses on the concept of gradational *inaccuracy* as something to be minimized. He makes the quite reasonable proposal that we assign true propositions the truth-value 1 and false ones the truth-value 0 and then define the inaccuracy of a degree of confidence c to P as some function of the absolute value of the difference between c and the truth-value of P.[13]

[12] The Dutch Book arguments for probabilism are synchronic. Diachronic Dutch Book arguments have been given for both simple conditionalization and Jeffrey conditionalization. See Skyrms (1990, chap. 5).

[13] Joyce (1998) allows for a family of functions as measures of inaccuracy, the most familiar of which is the Brier score. I explain the Brier score and illustrate its use as a measure of inaccuracy in the Appendix.

In an elegant proof, Joyce is able to show that whenever an agent's confidence assignments fail to satisfy the standard probability axioms, there is an alternative assignment that is more accurate in all classically possible worlds. Williams (2012) generalizes Joyce's result to apply to non-classical probabilities.

One problem with these results is that, even if an agent is quite sure there is a confidence assignment guaranteed to be more accurate than her own, there may be no rational way for a human being to figure out what it is. Consider again the example of Ian and the trillionth digit in the decimal expansion of π. Joyce's proof shows that, in all classically possible worlds, assigning confidence of 1.0 to the correct answer is guaranteed to be more accurate than a confidence assignment of .1 to each of the ten alternatives. But Joyce's proof cannot tell Ian which digit he should assign confidence of 1.0 to. So it cannot help Ian to improve the accuracy of his confidence assignment.

Although Joyce is a strict classical probabilist, the applicability of his concept of accuracy is not limited to confidence assignments that satisfy the probability axioms. Indeed, the special or general version of the MERF principle provides an example of a non-probabilist principle that minimizes expected inaccuracy (and maximizes expected accuracy), when expected inaccuracy is calculated on the basis of the MCS's determination of the relevant expected relative frequencies. If a confidence assignment is in equilibrium as defined by the special or general MERF principle and the MCS has opinions about the relevant relative frequencies of truth, any change will increase its expected inaccuracy (and thus decrease its expected accuracy).[14] For example, if Ian adopts a uniform confidence assignment to the ten possible digits in the sixth through 1.2 billionth digits in the decimal expansion of π and we interpret his MCS's estimates of the relevant relative frequencies of each digit in that expansion as a probability in the calculation of expected inaccuracy, the uniform confidence assignment minimizes expected inaccuracy (and thus maximizes expected accuracy) and any change would increase the expected inaccuracy of the assignment.

The final arguments for probabilism that I discuss are the potential calibration defenses. It is quite plausible that at least one goal of epistemically rational degrees of confidence is to satisfy the following external calibration condition, stated with an intuitive justification by Joyce: "What can it mean, after all, to assign degree of belief x to X if not to think something like 'Propositions like X are true about x proportion of the time?'(1998, p. 593) Van Fraassen (1983) and Shimony (1988) have argued that if external calibration in this sense is our goal, then *possible* external calibration should be a minimal condition of rationality. They argue that satisfying the standard probability axioms is necessary for the possibility of external calibration.

This is not true. Consider the example of Ian again. Ian's uniform confidence assignment to the ten possibilities for the sixth through the 1.2 trillionth digits in the decimal expansion of π is externally calibrated, because each digit will be correct almost exactly 10 % of the time.[15] As discussed above, Ian's uniform confidence assignment will also be in equilibrium as determined by the special MERF principle.

[14] For a proof, see the Appendix.

[15] Van Fraassen acknowledges the possibility of examples of this kind and then introduces a requirement that rules them out, because they are "irrational" (1983, p. 303). Why are they irrational? Van Fraassen's answer depends on their logically implying, by the laws of classical logic, an extension that is not well-calibrated; so the truth of the laws of classical logic are assumptions of his argument. Shimony rules out

Once it is acknowledged that the external calibration of a confidence assignment does not depend on its satisfying the standard probability axioms, it is possible to appreciate that the special or general MERF principle is an attempt to articulate the way in which external calibration should be understood to be a goal of epistemic rationality. External calibration per se is not the goal, because, as Joyce points out, it might be possible to attain perfect external calibration by assigning confidence of .5 to each proposition and its negation (1998, p. 595). The Narrower Reference Class Exception to the MERF principle has the effect of combining external calibration and accuracy into a single goal: to find the most accurate confidence assignments (closest to 1.0 or 0) that are externally calibrated. Both goals—external calibration and accuracy—can survive the move to non-probabilism. The example of Ian provides an example of a non-probabilist confidence assignment that both maximizes expected accuracy (relative to Ian's MCS's expectations for the relevant relative frequencies of truth) and is externally calibrated. In cases in which it is reasonable to translate degrees of confidence into betting quotients, any other betting quotients that Ian might employ would be irrational, though when offered a bet by someone more knowledgeable about the digits of π than he is, Ian would be best advised not to accept it.

5 Epistemic akrasia

One more example:

7. Epistemic akrasia. Richard acquires good evidence that his belief P is due solely to wish fulfillment and thus is unjustified. Other things being equal, Richard should stop believing P (Feldman 2005). *Why?*

There is a burgeoning literature on what has come to be referred to as *principles of epistemic akrasia*. They can be formulated as principles about justification: It is not rational to believe: P but I am not justified in believing P (Feldman 2005); or evidence: It is not rational to believe: P but my evidence does not support P (Horowitz 2014); or rational credence: It is not rational to hold credence of x in proposition P and to hold that the credence that is rational to place in P given my evidence is y and $x \neq y$ (Christensen 2010b).[16]

All instances of epistemic akratic principles are actually just special cases of metacognitive disequilibrium governed by the MERF principle. The MERF principle can subsume all of them—at least, to the extent that the other principles yield correct results in particular cases. For example, when Richard comes to believe that his belief P is unjustified because it is solely due to wish fulfillment, it is not necessary to invoke an epistemic akratic principle to explain why he should stop believing P. It is the fact that his MCS estimates the relative frequency of truth of beliefs solely due to wish fulfillment to be low (much less than .5) that explains why rationality requires a reduction in Richard's confidence in P to much less than .5. Once Richard has reduced

Footnote 15 continued
such "possibilities" from the outset, by simply assuming that the laws of classical logic determine what is possible (1988, p. 81).

[16] Though Christensen (2010b) does not fully endorse his principle, because of puzzles he discusses.

his confidence in P to something less than .5, it would be irrational for Richard to believe P.

Titelbaum (2015) provides his own explanation of the irrationality of believing P and believing that it is irrational to believe P. His explanation turns into a potential challenge to the main argument of this paper, at least as applied to epistemology. Before analyzing Titelbaum's argument, let me say something more about the operation of defeaters in reasoning.

6 Reliability defeaters

Pollock (1984) identifies two kinds of defeater: (1) rebutting defeaters and (2) undermining defeaters. I give an example of each.

Suppose Sheila has a friend Liz. Liz has often given Sheila rides to work in a Ford. So Sheila has lots of evidence, E_1, that make it rational for Sheila to believe that Liz owns a Ford. However, today when Liz picks Sheila up for work, she is driving a new Toyota. When Sheila comments on the new car, Liz tells her that yesterday she traded her Ford in for a new Toyota. Now Sheila has new evidence E_2 that defeats the rationality of her former belief that Liz owned a Ford, even though all of her prior evidence E_1 is still true. This is a rebutting defeater, because the new evidence is evidence against the truth of Sheila's belief that Liz owns a Ford.

The kind of defeater involved in metacognitive certainty-reducing reasoning is of a different kind, an *undercutting defeater*: John sees a chair that looks red and, on the basis of that perceptual evidence E_1, John rationally believes that it is red. If he then finds out that the chair is irradiated by red light and realizes that red light would make a white chair look red, that new evidence E_2 will reduce his confidence that the chair is red, not by rebutting his belief that the chair is red, but by undercutting it (Pollock 1984, p. 113). Metacognitive equilibrium reasoning involves a special kind of undercutting defeater, a *reliability defeater*, because the defeating evidence is evidence of the unreliability of the causal processes responsible for the undercut belief (114). If you have had the experience of realizing that you are dreaming while in the middle of a dream, then you have had the experience of acquiring a reliability defeater for your perceptual beliefs. You would have a reliability defeater for many of your current beliefs if all of a sudden you came to believe that you were dreaming right now.

A full theory of defeaters has an inductive structure, because there can be a defeater for a defeater; and so on. An *ultimate defeater* is one that is ultimately undefeated. In the simple examples I discuss, I will just take for granted that the defeaters that I discuss are ultimate defeaters, unless I specify otherwise.

7 Titelbaum's challenge

Titelbaum discusses an akratic principle that applies not only to theoretical rationality (rational belief) but also to practical rationality (rational intentions) (Titelbaum 2015, p. 261). Since I am focused on theoretical rationality, I will use brackets to show how to translate his general discussion into constraints on rational belief. Titelbaum argues that there is at least one kind of a priori belief, a special kind of epistemological belief, that is not liable to reliability defeat. I am going to use an example to explain the kind of belief involved. For the sake of comprehensibility, I will make simplifying assumptions, but nothing crucial will depend on them. Consider again the example of John and the chair. At the earlier time t_1, before John learns about the red light, I suppose that there is some overall, full specification of John's epistemic situation, including a specification of his relevant background beliefs BB_1 and his total evidence E_1, such that, all things considered, given BB_1 and E_1, rationality permits John to believe.

(RC) I am looking at a red chair

Titelbaum draws our attention to what I will call *situational epistemic beliefs*. Here is how he characterizes their content: "I will be asking whether, given an agent's current situation and taking into account every aspect of that situation pointing in whatever direction [i.e., the overall situation], it is all-things-considered rationally permissible [or required or forbidden] for her to adopt a particular combination of attitudes"(260) [my additions in brackets]. A *situational epistemic belief* has content of the following kind: It fully specifies an agent's epistemic situation (e.g., relevant background beliefs and total evidence) and asserts a relation between that specification and what it is all-things-considered rationally permissible (or required or forbidden) for the agent to believe in that situation (the agent's overall epistemic state) (259). Titelbaum limits his discussion to beliefs of this kind that are a priori (263).[17]

Consider again the example of the red chair. As explained above, at t_1 the following a priori situational epistemic belief would be true:

> $RP(BB_1, E_1, B(RC))$ = Given background beliefs BB_1 and total evidence E_1, it is rationally permitted to believe RC.

Titelbaum is going to argue that it is never rationally permissible to have a mistaken a priori situational epistemic belief. This is what he calls his *Fixed Point Thesis* (261), when it is applied to rational belief. Notice that the rational infallibility involved is not *situation-relative*, as it would be if Titelbaum were only claiming that it is not rationally possible to have a mistaken a priori situational epistemic belief *about one's own current epistemic situation*; rather the rational infallibility involved is *situation-invariant*, because Titelbaum is claiming that no a priori situational epistemic belief about *any situation* can be mistaken, *if it is rational* (263). Of course, once it is acknowledged that these beliefs are a kind of a priori belief, it is not surprising that the rational infallibility involved is situation-invariant. How could we plausibly explain the rational infallibility of any kind of a priori belief if, in some situations, an agent could be rationally mistaken about it (276)?

At this point, it would seem that we can just add Titelbaum to the list of philosophers who have claimed to have an a priori source of rational certainty in epistemology and use our estimate of the relative frequency of truth among such claims to justify assigning a very low degree of confidence to Titelbaum's claims of rational certainty.

[17] I set aside here the question of whether we have any beliefs that are purely a priori. I have my doubts about this, but I will try to show that even if there are purely a priori beliefs of the kind that Titelbaum employs in his account, his argument fails.

But Titelbaum's argument comes with a twist. He argues that anyone, including me, who accepts an epistemic akratic principle is committed to the rational infallibility of a priori situational epistemic beliefs (his Fixed Point Thesis); and, as a corollary, to the inapplicability of MERF-type reliability defeater reasoning to those beliefs.

Titelbaum sees no problem with the idea that empirical evidence could rationally undermine an empirical belief. But he argues that no empirical evidence could rationally undermine an a priori situational epistemic belief. To return to the example above, John's empirical belief RC—that the chair is red—can be defeated by empirical evidence, but Titelbaum would claim that, if rational, his situational epistemic belief RP(BB_1, E_1, B(RC))—that his background beliefs BB_1 and his total evidence E_1 (i.e., his overall epistemic situation) rationally permit believing that the chair is red—cannot be rationally undercut by empirical evidence. Call this Titelbaum's *no-reliability defeat-result* (for a priori situational epistemic beliefs). Here is how he states a situation-relative version of it:

> [E]very agent possesses a priori, propositional justification for true beliefs about the requirements of rationality in her current situation. An agent can reflect on her situation and come to recognize facts about what that situation rationally requires. Not only can this reflection justify her in believing those facts; the resulting justification *is also empirically indefeasible* (Titelbaum 2015, p. 276) [emphasis added].

Here Titelbaum is discussing the situation-relative rational infallibility of a priori situational epistemic beliefs. However, in the next paragraph, he makes the move from situation-relative to situation-invariant rational infallibility (276). So Titelbaum is committed to holding that all a priori situational epistemic beliefs are, when rational, infallible (his Fixed Point Thesis) and that, when rational, they are empirically indefeasible (the no-reliability defeat-result). Of course, if the no-reliability defeat-claim is true, it is irrational for someone with a true a priori situational epistemic belief (e.g., an epistemologist who truly believes that, his background beliefs and his evidence rationally permit or require belief in his epistemological theory) to give up that belief in the light of empirical evidence that other epistemologists disagree with him (288).

In the next section, I provide an example to show that the no-reliability defeat-result is false. In the following section, I explain why Titelbaum's positive argument for his Fixed Point Thesis fails.

8 The example of the rationality-impairing drugs

Is Titelbaum correct that there cannot be a rational reliability defeater for rational a priori situational epistemic beliefs? Consider this example:

Suppose that there are drugs that can be used to change a priori situational epistemic beliefs and to make enough other changes in the person's cognitive state to make her set of beliefs seem fully coherent from the inside. Mike has volunteered for what he believes to be a test of one of the drugs. It is actually a test of his a priori situational epistemic beliefs. When he is not under the influence of any of the rationality-impairing drugs, Mike is an epistemically rational agent all of whose beliefs are epistemically rational and none of which are epistemically irrational.

For generality, consider the possibility that there are n drugs, $D_1, \ldots, D_n$, (where D_n is a placebo) and n a priori situational epistemic claims, as follows: $RR(BB_i, E_i, B(P_i))$ = In an overall epistemic situation with relevant background beliefs BB_i and total relevant evidence E_i, rationality requires believing Pi.

Only $RR(BB_n, E_n, B(P_n))$ is true. All of the other $RR(BB_j, E_j, B(P_j))(1 \leq j < n)$ describe a powerful cognitive illusion that can be induced by the corresponding drug, D_j. I assume that the cognitive illusions only occur in ground-level cognitive processing; not in metacognititive reasoning.

I explain the experiment for n = 2. This simple experiment can be completed in a day. Mike is the experimental subject. The experiment takes place in a windowless room in which there is no evidence of the time of day. To keep track of the different stages of the experiment, it is useful to think of Mike as a union of two temporal selves—Mike_1 when he is under the influence of drug D_1 and Mike_2 when he is not under the influence of D_1:

6 am Mike_2 is hooked up to an intravenous drug delivery system and given D_1 in combination with a drug that induces dreamless sleep. Drug D_1 so alters Mike_2's ground-level cognitive processing to produce the powerful cognitive illusion that, given evidence E_1, P_1 is obviously true and P_2 is obviously false. D_1 also alters Mike_2's background beliefs to generate a set of background beliefs BB_1 such that to Mike_1 it seems obvious that $RR(BB_1, E_1, B(P_1))$ is true and $RR(BB_2, E_2, B(P_2))$ is false. $RR(BB_1, E_1, B(P_1))$ is also ideally coherent with Mike_1's background beliefs BB_1. As a consequence, when under the influence of the drug, there is no way from the inside for Mike_1 to use coherence considerations to cast doubt on $RR(BB_1, E_1, B(P_1))$, even though it is false. However, D_1 has no effect on Mike_1's metacognitive processes, because the cognitive illusion it produces is only involved in ground-level, not metacognitive, processing.

For the example to be persuasive, you have to imagine that drug D_1 will lead Mike_1 to *seem* to be able to justify his mistaken a priori situational epistemic belief $RR(BB_1, E_1, B(P_1))$ with what seem, from the inside, to be rationally compelling reasons, seemingly just as rationally compelling as the true reasons that support Mike_2's rational rejection of $RR(BB_1, E_1, B(P_1))$ when he is not under the influence of the rationality-impairing drug D_1.[18] It would be too much of a diversion for me to undertake to specify the example in full enough detail to illustrate all the symmetries in reasoning that would be required. My goal is to provide just enough detail that you

[18] This requirement places some limits on the content of P_2. For example, I doubt that, even in theory, there could be a drug that would make it possible to coherently believe that our current evidence rationally requires belief in any alternative to the belief "I exist." I think we have enough evidence of powerful cognitive illusions in human beings to be quite confident that there are some values of P_2 for which, in theory, there could be drugs with the effects I describe in the text. If this is correct, then perhaps *some*, but not *all*, rational a priori situational epistemic judgments are infallible. I can allow for this possibility and still be a thoroughgoing fallibilist even about rational a priori situational epistemic beliefs. Even if some sub-class of those beliefs are infallible, it would not be rational for us to be certain of any one of them, if, as I believe, there is no infallible way of drawing the line between those that are infallible and those that are not.

can use your imagination to fill in missing details in a way to that would establish the required symmetries.

8 am Mike$_1$ is awakened. Under the influence of drug D$_1$, Mike$_1$ is first asked to identify his relevant background beliefs and total evidence, which he correctly identifies as BB$_1$ and E$_1$. Then he is asked two questions: (1) Which, if any, of the following are true: P$_1$ or P$_2$? To which he replies: P$_1$ (which, by hypothesis, is false). (2) Which, if any, of the following are true: RR(BB$_1$, E$_1$, B(P$_1$)) or RR(BB$_2$, E$_2$, B(P$_2$))? To which he replies: RR(BB$_1$, E$_1$, B(P$_1$)) (which, by hypothesis, is also false). When Mike$_1$ is asked to explain his answers, he does so in a way that seems coherent to him given his background beliefs. Then he is given the drug that induces dreamless sleep, an amnesiac that erases all memory of his 8 am awakening, an antidote to D$_1$, and D$_2$, a placebo.

10 am Mike$_2$ is awakened. Not under the influence of any rationality-impairing drug, he correctly identifies his background beliefs BB$_2$ and total evidence E$_2$. He is again asked the same two questions: (1) Which, if any, of the following are true: P$_1$ or P$_2$? To which he replies: P$_2$ (which, by hypothesis, is true). (2) Which, if any, of the following are true: RR(BB$_1$, E$_1$, B(P$_1$)) or RR(BB$_2$, E$_2$, B(P$_2$))? To which he replies: RR(BB$_2$, E$_2$, B(P$_2$)) (which, by hypothesis, is also true). When he is asked to explain his answers, he does so in a way that seems coherent to him given his background beliefs. Then he is given a drug that induces dreamless sleep, an amnesiac that erases all memory of his 10 am awakening, and drug D$_1$.

Noon Mike$_1$ is awakened. The experimental protocol from 8 am is repeated, with very much the same results. Then Mike$_1$ is debriefed by the experimenters. The experimental protocol for the entire experiment is explained to Mike$_1$, including the information that D$_1$ affects only ground-level, not metacognitive level processing. However, he is not given any information that would help him to determine when he was under the influence of D$_1$ and when not. Then he is shown videotapes of the two earlier sessions. He is reassured to see one of his earlier selves make what seems to him the obviously correct judgments P$_1$ and RR(BB$_1$, E$_1$, B(P$_1$)), but he is shocked to see another of his earlier selves make what seems to him the clearly irrational judgments P$_2$ and RR(BB$_2$, E$_2$, B(P$_2$)).

1 pm (a continuation of Mike$_1$'s noon wakening) After viewing the videotapes, and still under the influence of drug D$_1$, Mike$_1$ is asked the same two questions again. I postpone discussion of his answers. Then he is given a drug that induces dreamless sleep, an amnesiac that erases all memory of his noon awakening, an antidote to D$_1$, and drug D$_2$ (the placebo).

2 pm Mike$_2$ is awakened. The same protocol is followed as at the noon awakening. After initially answering P$_2$ and RR(BB$_2$, E$_2$, B(P$_2$)), Mike$_2$ is debriefed and then shown the two videotapes. Again he is shocked, but this time the shock is that one of his earlier selves made what seem to him to be the clearly irrational judgments P$_1$ and RR(BB$_1$, E$_1$, B(P$_1$)). After listening to his earlier self defend P$_1$ and RR(BB$_1$, E$_1$, B(P$_1$)), he comes to the realization that when that self views the videotape of his earlier self's assertions of P$_2$ and RR(BB$_2$, E$_2$, B(P$_2$)), that self will regard those opinions to be as irrational as *he* now regards the opinions P$_1$ and $RR(BB_1, E_1, B(P_1))$.

3 pm (a continuation of Mike$_2$'s 2 pm wakening) After viewing the videotapes, and still free of the influence of D_1, Mike$_2$ is again asked the same two questions again. If he is rational, how will he answer?

If Titelbaum is correct, at 3 pm, after he has been debriefed and he has viewed the two videos, rationality requires that Mike$_2$ continue to believe $RR(BB_2, E_2, B(P_2))$. This is because Titelbaum denies that a rational a priori situational epistemic belief can be empirically defeated. Since, by hypothesis, Mike$_2$'s earlier situational epistemic belief $RR(BB_2, E_2, B(P_2))$ at 2 pm (as well as at 10 am) is rational (and true) and Mike$_2$ has only been given empirical information since being awakened at 2 pm, according to Titelbaum, rationality requires that, if he has any opinion at all, Mike$_2$'s a priori situational epistemic judgment at 3 pm agree with his earlier judgment at 2 pm (as well as at 10 am) that $RR(BB_2, E_2, B(P_2))$.[19]

I think this is a mistake. I think that at 3 pm rationality requires Mike$_2$ not to believe either of the alternatives $RR(BB_1, E_1, B(P_1))$ or $RR(BB_2, E_2, B(P_2))$, and that if Mike$_2$ is rational at 3 pm, he will recognize that rationality requires that he not believe either of them. This is because I suppose that, at 3 pm, when Mike$_2$'s MCS evaluates the reliability of his ground-level processing, he finds that he has good reason to trust what the experimenters have told him in the debriefing, which is at least partly confirmed by the two videos. Given the information that the rationality-impairing drug does not affect metacognitive processing, when his MCS estimates the expected relative frequency of truth of his relevant situational epistemic judgments, he will have no basis for judging either one of his earlier selves' beliefs to be more reliable than the other. His MCS will rationally believe that one of those judgments is due to a powerful cognitive illusion, reinforced by compelling background beliefs, but he won't have any way to tell which one it is. His MCS will categorize the causal processes responsible for each of the two earlier judgments, $RR(BB_1, E_1, B(P_1))$ and $RR(BB_2, E_2, B(P_2))$, disjunctively as either due to ordinary reasoning or due to drug-influenced reasoning and will estimate the expected relative frequency of truths in those judgments to be no higher than .5; this will throw his confidence assignments into metacognitive disequilibrium. Equilibrium will be restored by reducing his confidence in each judgment to .5 or less, so that he won't believe either of them.

I do not expect this thought experiment to change Titelbaum's mind. So I suspect that at this point we would reach an impasse. Titelbaum thinks that the only way that I could rationally reconcile myself to the impasse would be for me to commit myself to the existence of genuine rational dilemmas (290–291). I think there is another way out, as I explain shortly.

[19] It might seem that Titelbaum could save his rational infallibility claim by holding that, after the debriefing, rationality does not permit Mike to believe anything about the situation in which an agent has relevant background beliefs BB2 and total evidence E2; not even that very statement. I think this is a mistaken conclusion to draw about the example described in the text, as I explain in the next paragraph. But even if rationality did not permit Mike to believe anything about his situation after the debriefing, the example would still be an example of empirical defeat of an a priori situational epistemic belief. Titelbaum does consider the possibility that his opponent might be stuck in a kind of rational dilemma (290), though he himself does not endorse this position as rational. I discuss what Titelbaum takes to be the rational dilemma shortly.

Some readers might think that rationality allows for us to give special weight to our own opinions; so that at 3 pm, after viewing the videotapes for the second time, because Mike_2 is not under the influence of any rationality-impairing drug, rationality would at least permit him to endorse the a priori situational epistemic belief $\mathrm{RR}(\mathrm{BB}_2, \mathrm{E}_2, \mathrm{B}(\mathrm{P}_2))$. I think this is a mistake even in this case, where n = 2; but I would expect this move to become progressively more implausible as n increases. For example, when n = 10, there will be ten different drugs, $\mathrm{Di}(1 \le \mathrm{i} \le 10)$, the first nine of which are rationality-impairing and the tenth, D_{10}, is a placebo, and ten different a priori situational epistemic beliefs, $\mathrm{RR}(\mathrm{BB}_\mathrm{i}, \mathrm{E}_\mathrm{i}, \mathrm{B}(\mathrm{P}_\mathrm{i}))(1 \le \mathrm{i} \le 10)$, only one of which, $\mathrm{RR}(\mathrm{BB}_{10}, \mathrm{E}_{10}, \mathrm{B}(\mathrm{P}_{10}))$, is true; the other nine would be the result of powerful cognitive illusions.

To visualize this case, it is helpful to divide Mike himself into ten temporal stages: $\mathrm{Mike}_\mathrm{i}(1 \le \mathrm{i} \le 10)$, are the temporal stages of Mike, the first nine of whom are under the influence of one of the nine rationality-impairing drugs, and Mike_{10} whose rationality is not impaired, because drug D_{10} is a placebo. In the final stage of the experiment, each Mike_i will view videotapes of the ten earlier stages of the experiment, and each Mike_i will initially agree with the earlier Mike_i's judgment and initially judge that the other Mike_k's $(1 \le \mathrm{k} \le 10, \mathrm{k} \ne \mathrm{i})$ are subject to a powerful cognitive illusion. But once the Mike_i recognize the symmetry of their situations, I hold that, for all ten of them, their MCSs should rationally revise their confidence in the $\mathrm{RR}(\mathrm{BB}_\mathrm{i}, \mathrm{E}_\mathrm{i}, \mathrm{B}(\mathrm{P}_\mathrm{i}))(1 \le \mathrm{i} \le 10)$, to make them equal the expected relative frequency of truth of .1.

To dramatize the difference between Titelbaum and me when n = 10, suppose that after viewing the videotapes, each of the Mike_i is offered an even money $100 bet on $\mathrm{RR}(\mathrm{BB}_\mathrm{i}, \mathrm{E}_\mathrm{i}, \mathrm{B}(\mathrm{P}_\mathrm{i}))$—that is, on the a priori situational epistemic belief that seemed obviously true to them before they were debriefed and viewed the videotapes; and suppose that they evaluate bets in the usual way. Suppose, also, that if Mike had been offered the bet before the experiment, he would have immediately taken it, because $\mathrm{RR}(\mathrm{BB}_{10}, \mathrm{E}_{10}, \mathrm{B}(\mathrm{P}_{10}))$ seems so obviously true to him and the other alternatives seem so obviously false. On my account, since none of the drugs affects metacognitive processing, each of the Mike_i will be metacognitively rational enough to assign confidence of .1 to each of the ten alternatives $\mathrm{RR}(\mathrm{BB}_\mathrm{j}, \mathrm{E}_\mathrm{j}, \mathrm{B}(\mathrm{P}_\mathrm{j}))(1 \le \mathrm{j} \le 10)$, including $\mathrm{RR}(\mathrm{BB}_\mathrm{i}, \mathrm{E}_\mathrm{i}, \mathrm{B}(\mathrm{P}_\mathrm{i}))$; no Mike_i will accept the proffered bet and Mike will lose no money. On Titelbaum's account, it would be rational for Mike_{10} to accept the bet on $\mathrm{RR}(\mathrm{BB}_{10}, \mathrm{E}_{10}, \mathrm{B}(\mathrm{P}_{10}))$. As he is making that bet, Mike_{10} will realize that, since none of the drugs affects metacognitive processing, each of the nine other Mike_i's will reason the same way that he has and each will accept the bet on $\mathrm{RR}(\mathrm{BB}_\mathrm{i}, \mathrm{E}_\mathrm{i}, \mathrm{B}(\mathrm{P}_\mathrm{i}))$; and each of the other Mike_i's will recognize this also. As a result, Mike_{10} (and each of the other Mike_i's) will realize that they will jointly enter into a combination of bets on which, they all agree, Mike (as the union of the Mike_i) will sustain a sure loss of $800.

This betting scenario is not a decisive objection to Titelbaum's view, because it is always open to him to reply that Mike's rational self is not responsible for the actions of his irrational selves (cf. Titelbaum 2015, p. 287). It is puzzling, however, to think that each of the Mike_i might say exactly the same thing to explain why their bet is not irrational, even though none of them would have any basis for distinguishing *his belief*

$RR(BB_i, E_i, B(P_i))$ as the rational one and the corresponding beliefs of the other nine as irrational.

Although Titelbaum does not consider any example like the rationality-impairing drug example, he does briefly consider the epistemic situation of an opponent who, having found Titelbaum's positive argument for the Fixed Point Thesis to be compelling, also finds it compelling that a priori situational epistemic beliefs can be subject to rational reliability defeat. He thinks that the only rational way out for such an opponent is to acknowledge the existence of genuine rational dilemmas, situations in which "*there are no rationally flawless options*"(290–291; emphasis in original). As Titelbaum correctly insists, even if one grants that there are genuine rational dilemmas, this is not a reason to deny the Fixed Point Thesis (291).

However, Titelbaum is mistaken to think that his opponent's only way out is to acknowledge the existence of genuine rational dilemmas. There is another way out: To show that Titelbaum's positive argument for the Fixed Point Thesis fails.

9 Titelbaum's positive argument for the fixed point thesis

Titelbaum's argument for the rational infallibility of a priori situational epistemic beliefs (his Fixed Point Thesis) is an explanationist argument. It starts from a version of the Akratic Principle, one that applies to theoretical reason (rational beliefs) and practical reason (rational intentions). I continue to use brackets to illustrate how Titelbaum's principle applies to rational belief:

> (*Akratic Principle*) No situation rationally permits any overall state containing both the attitude A [e.g., the belief P] and the belief that attitude A [i.e., believing P] is rationally forbidden in one's current situation (Titelbaum 2015, p. 261).

Titelbaum argues that the rational infallibility of a priori situational epistemic beliefs is part of the "most obvious" explanation of the Akratic Principle (276). Again, I use brackets to illustrate how his statement of the conclusion of his explanationist argument applies to rational belief:

> How is the justificatory map arranged such that one is never all-things-considered justified in both an attitude A [e.g. the belief P] and the belief that A [i.e., believing P] is rationally forbidden in one's current situation? The most obvious answer is that every agent possesses a priori, propositional justification for true beliefs about the requirements of rationality in her current situation. An agent can reflect on her situation and come to recognize facts about what that situation rationally requires (276).

This is Titelbaum's statement of the conclusion of his explanationist argument for the situation-relative rational infallibility of a priori situational epistemic beliefs. In the next paragraph he asserts that every plausible account of the truth of the Akratic Principle requires not only situation-relative rational infallibility, but also situation-invariant rational infallibility of those beliefs (his Fixed Point Thesis).

I am going to challenge Titelbaum's argument for his Fixed Point Thesis in three ways:

(1) By showing that his argument for the situation-relative rational infallibility of a priori situational epistemic beliefs fails. The argument fails because it is based on a false presupposition about the nature of reasoning. The false presupposition is most clearly revealed in Titelbaum's discussion of reasoning about testimony, but his discussions of reasoning about perception and logical reasoning show that it is implicated in his understanding of those kinds of reasoning also (272–273). The false presupposition is that there are exceptionless inferential principles of reasoning about testimony or perception or of logical reasoning. I believe that this is a mistake. There are no such principles because all such reasoning is holistic.

(2) By showing that even if the Fixed Point Thesis were true, it would not explain the Akratic Principle, as that principle is most naturally interpreted and as it has been advocated and defended in the literature (e.g., Horowitz 2014), because the Fixed Point Thesis does not rule out all exceptions to the Akratic Principle. The Fixed Point Thesis can explain only a limited version of the Akratic Principle, not the fully general version. The only remotely plausible way that the Fixed Point Thesis could explain the truth of the fully general version of the Akratic Principle—that is, the principle as it has been advocated and defended in the literature—would require even more implausible rational infallibility claims, claims concerning the rational infallibility of beliefs about our own mental states.

(3) By providing an alternative explanation of the fully general Akratic Principle. It is a better explanation than Titelbaum's, in part, because it explains *all* of the instances of the general Akratic principle, not merely the instances of a limited version of it and, in part, because it does not require any kind of infallibility.[20] This alternative explanation of the Akratic Principle is one that subsumes it under the MERF Principle.

I discuss each of these three claims in order.

9.1 Titelbaum's mistaken presupposition about reasoning about testimony (and other kinds of reasoning)

The first problem with Titelbaum's argument for the situation-relative rational infallibility of a priori situational epistemic beliefs is that it presupposes a mistaken presupposition about reasoning. The mistaken presupposition is that reasoning about testimony can be reconstructed in terms of exceptionless principles of inference (e.g., 270–277).

Titelbaum notes that his opponent, who believes in the potential for undermining defeaters (e.g., by testimony), will need some principles to explain when it is rational to accept the sources of those defeaters (e.g., testimony). So Titelbaum offers his opponent some sample principles and finds problems with them, which he fixes by including a proviso that secures the rational infallibility of a priori situational epistemic beliefs.

[20] The MERF principle explanation has other explanatory advantages over Titelbaum's account, but it is not necessary to address them here.

But Titelbaum's argument implicitly assumes that there is some true exceptionless rule of inference of this form:

> (*Testimonial Inference Schema*) If an agent's situation includes testimony that P, the agent is rationally permitted and required to believe P, unless Q (cf. Titelbaum 2015, pp. 270–275).

Titelbaum argues that no principle fitting this schema will be adequate to explain rational belief about testimony unless Q includes a condition that involves a commitment to the rational infallibility of a priori situational epistemic beliefs; otherwise, there will exceptions (Titelbaum 2015, pp. 274–278). Titelbaum's mistake is to think that there is a non-trivial exceptionless rule of inference fitting the form of the Testimonial Inference Schema. The reason is that all (or almost all) reasoning, including reasoning about testimony, is holistic and holistic reasoning does not fit the schema above, except in a trivial way.[21]

In holistic reasoning the contents of one's entire set of beliefs, as well as other sources of input, determine what it is rational to believe, because what it is rational to believe is ultimately the product of a largely unconscious all-things-considered determination of what makes the most sense. No one has ever come close to articulating exceptionless principles of rational belief, in part because those principles would take as input a person's entire set of beliefs as well as any other sources of input (e.g., perception) and would output an entire set of beliefs, the set of beliefs that make the most sense, given the input. Since no one has ever come close to articulating her entire set of beliefs, we have very little idea of what the holistic principles of reasoning are, other than to say that they determine the overall coherence of one's set of beliefs.[22]

Holistic reasoning is nonmonotonic, it can add beliefs and it can subtract beliefs. In contrast, the schema for testimonial inference above reads like a rule of inference with an application condition. Rules of inference are monotonic, so they can never be adequate to understanding holistic reasoning.

It is sometimes useful to use inference rules for reconstructing reasoning about testimony or perception or for reconstructing logical reasoning, if they are used as rules of thumb—that is, on the understanding that they only hold other things being equal—in other words, that they have exceptions. I believe that this is true even of logical reasoning. Not even logical reasoning is monotonic, because it is not always rational to accept the logical consequences of our rational beliefs. If the conclusion of the reasoning does not cohere well with other rational beliefs (the holistic element),

[21] The "almost all" qualification is meant to leave it open that there might be some very narrow kinds of reasoning that are not subject to defeat—for example, Descartes' inferences from propositions about what he is thinking to the propositions that he exists. I think that even this kind of reasoning may be subject to rational defeat, but it is beyond the scope of this paper to discuss such issues. My claims here are limited to inferential principles for reasoning about testimony and perception, and for logical reasoning. Also, for ease of exposition, I follow Titelbaum in considering principles of reasoning that apply to beliefs, even though I believe that the most general principles of reasoning apply to degrees of belief.

[22] We have very little understanding of the standards of coherence. My earlier discussion of fallibility about the laws of logic implies that not even the laws of logic are standards of rational coherence. The kind of coherence involved is explanatory coherence, but no one has a very good understanding of what that is.

it may make more sense overall to reject one of the premises rather than to accept the conclusion (the nonmonotonic element).[23]

The only way to turn the testimonial inference schema above into an exceptionless rule of reasoning would be to substitute for Q: *all-things-considered you are not permitted and required to believe P*. This would make the schema exceptionless, but trivially true.

In the history of epistemology, there are many examples of inferential models of reasoning that implied some kind of infallibility. I believe that we should reject those models of reasoning in favor of a holistic, coherence theory of reasoning.[24] If we reject the schema that Titelbaum offers us for a rule of testimonial inference, his argument for the infallibility of rational a priori situational epistemic beliefs collapses.

9.2 Even if the Fixed Point Thesis were true, it would not rule out all exceptions to the Akratic Principle, as that principle has been articulated in the literature

The Akratic Principle, as it is usually understood, asserts that the following combination can *never* be rational: believing P and believing that rationality requires not believing P. Although Titelbaum asserts that the Fixed Point Thesis is the "most obvious" explanation of the Akratic Principle (276), he is mistaken, because even if the Fixed Point Thesis were true, it would not rule out all exceptions to the Akratic Principle, as it is usually understood.

To see this, note that, as Titelbaum himself recognizes, no a priori situational epistemic belief by itself implies anything about what it is rational for an agent to believe that she is rationally required to believe or not to believe in her current epistemic situation, which is what the Akratic Principle constrains. Consider the epistemic situation of a rational agent Michaela. Michaela, we suppose, has infallible a priori situational epistemic beliefs. How can Michaela use those beliefs to draw conclusions about what she is rationally required to believe or not to believe in her current epistemic situation? As Titelbaum himself acknowledges, she must augment those a priori beliefs with a posteriori beliefs about her current situation and her overall epistemic state and Titelbaum acknowledges that an agent can have mistaken a posteriori beliefs of these kinds (263). To simplify my argument, suppose that Michaela's current situation is fully specified by her current background beliefs BB and evidence E. If it is possible for a rational agent to have mistaken beliefs about her background beliefs or her evidence, then even if the Fixed Point Thesis is true, it will not prevent Michaela from failing to satisfy the Akratic Principle, as it is usually understood.

[23] I should mention that there are other alternatives that could be more rational than to accept the conclusion of logical reasoning: to give up the belief that the premises deductively imply the conclusion, either because one made a mistake about which rules are deductively valid or because one made a mistake in thinking that the inference in question was an instance of those rules.

[24] I should add that it is possible to believe that all reasoning is a kind of coherence reasoning without being a coherence theorist of rationality, because one can allow for rational input, so long as the input beliefs themselves are thought of as rationally defeasible.

To see this, suppose that Michaela rationally but mistakenly believes that her background beliefs are BB' when in fact they are BB. Suppose also that she correctly identifies her current evidence E. Michaela has infallible knowledge that given background beliefs BB and evidence E, rationality requires believing P; and that given background beliefs BB' and evidence E, rationality requires not believing P. Because Michaela is rational and does in fact have background beliefs BB and evidence E, rationality requires Michaela to believe P, and she does so. However, because Michaela rationally but mistakenly believes that her background beliefs are BB', she mistakenly concludes that in her current situation, rationality requires *not* believing P. Thus, Michaela satisfies Titelbaum's Fixed Point constraint, but she violates the Akratic Principle, as it is generally understood in the literature (e.g. Horowitz 2014), because she believes P and she believes that rationality requires not believing P in her current situation.

Don't misunderstand me. Michaela's overall epistemic state does violate the general version of the Akratic Principle and it is irrational. However, the Fixed Point Thesis cannot explain what is irrational about Michaela's overall epistemic state. Shortly, I explain why her overall epistemic state is irrational.

Titelbaum seems to be aware that there are possibilities of this kind. Here is what he says:

> Rational evaluations in such cases are subtle and complex. The Akratic Principle might seem to indict the agent's overall state in all these cases, and I don't want to be committed to that. I have tried to formulate the principle carefully so as to apply only when an agent has the belief that her current situation, *described as her current situation*, rationally forbids a particular attitude. But that formulation may not handle all complications involving multiple descriptions of situations, and it certainly doesn't handle failures of state luminosity. Frankly, the best response to these objections is that while they are important, they are tangential to our main concerns here (262) [emphasis in the original].

But the example of Michaela is not tangential to our main concerns, because what Titelbaum seems to be admitting here is that his Fixed Point Thesis cannot explain the Akratic Principle as it is generally understood, including the most natural interpretation of his own statement of it, which is that it is *never* rational for an agent to believe P and to believe that rationality requires not believing P—even in cases involving failures of *state luminosity*—that is, in cases in which the agent has mistaken beliefs about her current situation (e.g., about her background beliefs or her evidence). Here Titelbaum is implicitly acknowledging that his Fixed Point Thesis only explains a limited version of the Akratic Principle, one that only applies to agents who correctly ascertain their current situation:

> (*Limited Akratic Principle*) No situation rationally permits any overall state containing both the attitude A [e.g., the belief P] and the belief that attitude A [i.e., believing P] is rationally forbidden in one's current situation, *if the agent correctly ascertains her current situation (including her background beliefs and evidence).*

If Titelbaum had presented his argument for the Fixed Point Thesis as an argument that his Fixed Point Thesis explains the Limited Akratic Principle, the obvious response would have been: Why the limitation? It certainly seems to be irrational to believe P and believe that rationality requires not believing P *in any situation*, even in situations, such as Michaela's, in which the agent is mistaken about her background beliefs or evidence.

The only remotely plausible way that Titelbaum's Fixed Point Thesis could explain the fully general Akratic Principle would be to augment it with a *luminosity thesis* that Titelbaum explicitly rejects, the thesis that "all aspects of situations are luminous to the agents in those situations"(262)—that is, that a posteriori beliefs about all aspects of one's current epistemic situation are rationally infallible. Given that one's set of background beliefs is clearly a part of one's epistemic situation and that no one has ever written down an exhaustive list of her background beliefs—indeed, I doubt that anyone ever could—the luminosity thesis is quite implausible. But unless the luminosity thesis is true, Titelbaum's own Fixed Point Thesis cannot explain the general version of the Akratic Principle. The most that Titelbaum can claim for his Fixed Point Thesis is that it can explain the Limited Akratic Principle.

9.3 There is an alternative explanation of the fully general Akratic Principle that does not require any kind of infallibility

The alternative explanation is simple: The fully general Akratic Principle is true—that is, it is never rational to believe P and to believe that rationality requires not believing P—because the belief that rationality requires not believing P is a reliability defeater for the belief P; where the fact that it is a reliability defeater for the belief P is explained by the operation of the MERF principle (cf. Talbott 1990, Preface 44). Notice that an explanation of reliability defeat in terms of the MERF principle will explain why no combination of the belief P with the belief that rationality requires not believing P will be metacognitively rational—that is to say, no combination of a rational *or irrational* belief P with a rational *or irrational* belief that rationality requires not believing P will survive the operation of the MERF principle.

Consider how the MERF principle would apply to Michaela's situation, discussed above. When Michaela comes to believe P and to believe that rationality requires her not to believe P, her MCS will classify her belief P as something that she is rationally required not to believe and then estimate the expected relative frequency of truths in propositions that she is rationally required not to believe. Michaela's estimate of this relative frequency of truth will be very low, much lower than .5. Since to believe P, her ground-level confidence assignment to P must be greater than .5, her confidence assignment to P will be in metacognitive disequilibrium, as defined by the MERF principle. Any restoration of metacognitive equilibrium will bring Michaela into compliance with the Akratic Principle. Typically, the restoration of equilibrium will occur in one of two ways: either by reducing her confidence in P to equal her MCS's estimate of the expected relative frequency of truths in propositions that she is rationally required not to believe, which will result in her no longer believing P; or by reevaluating her confidence assignment to the proposition that rationality requires her not to believe P so as ultimately to give up that belief. In either case, metacognitive rationality prevents her from believing both P and that rationality requires her not to believe P.[25]

It is time to take stock. In the previous section, I gave an example to show that Titelbaum's no-reliability defeat-claim is false. In this section, I have argued that his positive argument for the Fixed Point Thesis fails, because it is based on a mistaken model of reasoning; because even if the Fixed Point Thesis were true it would not explain the general version of the Akratic Principle that has been advocated and defended in the literature, the version that one obtains from the most natural reading of Titelbaum's own statement of it; and because there is an alternative explanation of the general version of the Akratic Principle that does not require any kind of infallibility: It is a simple consequence of the operation of the MERF principle in rational metacognition.

Let me close this discussion of Titelbaum by noting that it would be a mistake to think that my argument implies that rationality requires abandoning an epistemological view whenever anyone rationally disagrees with it or when a majority of epistemologists disagree with it. In any area of inquiry, it is valuable for different inquirers to *accept*, for the purpose of advocating them, views that they cannot rationally believe to be true. This is because, on my view, rationality is a kind of implicit sensitivity to the requirements of rationality. The social–historical process of epistemology itself would not be rational, unless individual epistemologists were sensitive to different aspects of rationality. If everyone who had a minority view in epistemology abandoned it and adopted the majority view, progress in epistemology would slow and, perhaps, even come to a halt. Only if all us do our best to articulate the aspects of rationality that we are sensitive to does epistemology have any realistic chance of progressing toward the truth. So I don't regard it as bad news that Titelbaum will almost surely disagree with me on the example of the rationality-impairing drugs. We need rational disagreement in epistemology, even rational disagreement on what there can be rational disagreement about. When we are engaged in doing epistemology, we need to make the best case we can for our own position. But when we step back and evaluate our own contributions, our metacognitive selves should make us extremely modest about the confidence we place in their truth.

[25] My account relies on there being some sort of conceptual connection between the reliability-relevant property of my being something that I am rationally required not to believe and the property of having a low expected relative frequency of truth. If we translate Pollock's use of *being reliable* into my terms as *having a reliability-relevant classification with a high estimated relative frequency of truth* and equate Pollock's use of *being a justified belief* with *being an (epistemically) rational belief*, then Pollock (1984) offers an insightful discussion of this conceptual connection. See also Talbott (1990). Of course, when I assert that something is a conceptual truth, I do not mean to imply that it is rational to be certain of it. Given the history of mistakes in what were thought to be conceptual truths, the MERF principle can easily explain why it is not rational for us to be certain of propositions that we take to be conceptual truths. It is just such considerations that led BonJour to become a fallibilist about all a priori justification (BonJour 1998, chap. 4).

10 What is the relation between probability and epistemically rational degrees of confidence?

This paper has been an extended argument (though only one part of a much longer extended argument) for a non-probabilist theory of rational degrees of confidence. Along the way I have argued against some of the presuppositions of any probabilist epistemology—for example, by arguing that rational empirical certainty-reducing reasoning could lead us to reduce our certainty in the laws of logic or that it could lead us to be certain of nothing. But there is an explanatory puzzle for my position. Probability seems to play an important role in my account. For example, I believe that the goal of rationality is to produce externally calibrated confidence assignments that are as accurate as possible. A confidence assignment that achieved the goal would define a probability assignment, because actual relative frequencies of truth in finite sets always satisfy the probability axioms. In addition, the operation of my MERF principle actually seems to be a consequence of a theorem of personal probability theory (from de Finetti [1937]), which Jeffrey calls *de Finetti's law of small numbers*. Here is one way that Jeffrey states the law:

> *De Finetti's Law of Small Numbers: Your estimate of the relative frequency of truths among the propositions* A_1*, ...,An must equal the average of the probabilities you attribute to them* (Jeffrey 1986, p. 54).

This law states that your estimate of the relative frequency of truths in a set S must equal the average of your confidence assignments to the members of S. Notice that the MERF principle is a direct consequence of the law of small numbers, if we just substitute "confidence assignments" for "probabilities" in Jeffrey's statement of the law. So we must wonder: Is it just a coincidence that probability plays such a large role in my non-probabilist theory?

Jeffrey (1986) would argue that it is no coincidence. He would argue that the fact that, on my own view, the goal of rationality is to generate a probability function is a good reason for thinking that probabilism is true.[26]

If the fact that the MERF principle is implied by the law of small numbers is not merely a coincidence, how could a non-probabilist like me explain why it is not? I agree that my account of epistemic rationality is implicitly committed to some kind of probability, but not to probabilism. What alternative is there? To answer that question, let me say something more about the nature of epistemic rationality.

To begin with, notice that the probability axioms place a priori content constraints on what we could rationally believe, because they require a priori that some propositions be assigned confidence of 1.0 purely on the basis of their logical structure. If we think of theories of rationality as theories of what it is to be a good learner, all probabilist theories set content limits on what a rational agent could learn. For example, a classical probabilist agent could not learn that there were exceptions to the law of excluded middle. When I raise this objection to probabilists, they typically reply, that is no

[26] Thanks to an anonymous referee for making a version of this argument.

problem, because there are non-classical versions of probabilism. If some non-classical logic is true, we will just adopt a non-classical version of probabilism.

But this response misses something important. If no probabilist agent could *learn* which logical laws are true, then we need some other kind of agent to learn which logical laws are true so that *that* agent can tell us what propositions the probabilist agent should assign probability of 1.0 to. One thing we know about *that* agent is that whatever kind of rationality it has, if it can rationally correct mistaken beliefs about the logical laws, as it must if it is able to *learn* which logical laws are true, it cannot be a probabilist agent. Now my suggestion is that a theory of human rationality is a theory of *that* kind of non-probabilist agent. And I don't see why it would make sense to regard that kind of rationality as 'non-ideal', if it is necessary for someone to have that kind of rationality in order for us to know how to define the 'idealized' probabilist agent.

In any case, my thought is that a truly rational agent would be able to reconsider any of her beliefs in light of evidence that they were false, even her beliefs about the laws of logic; and a truly rational agent would have to be open to the possibility that any (or almost any) belief—including beliefs about the laws of logic—could be subject to rational empirical reliability defeat. This does not imply that there would be no default beliefs built into her cognitive system. No human being could survive if there were not reasonably accurate default beliefs built into her cognitive system. We know that there are such default beliefs in our cognitive systems from the evidence of visual illusions and cognitive illusions (e.g., that space is Euclidean). Such default beliefs only need be true enough that we can survive long enough to get evidence that they are not true. But to be rational, we have to be able to use certainty-reducing reasoning to correct those illusions, because, for example, before we study physics, it really does seem to us, as it did to Kant, that it is certainly true that the universe is Euclidean.

Probabilist epistemology has a further learning limitation built into it. It is an immodest epistemology, so it could never explain how someone could rationally give up at least some of her probabilist epistemological beliefs. I believe there is a single theory of epistemic rationality, understood as a theory of what it is to be a good learner, and it applies to every domain, including logic and epistemology. One indication that such a theory is possible is that when, given our evidence, the MERF principle is applied to itself, it is not immodest. Indeed, given our evidence of the history of mistakes in epistemology, when applied to itself, the MERF principle is extremely modest; it does not even imply that we should believe it, because it cannot support confidence anywhere near .5 in substantive epistemological principles, including itself. In fact it can be used to explain what is almost certainly true, that it is rational to believe that there is further room for improvement in the formulation of it.

If a theory of the rationality of belief and degrees of confidence is a theory of what it is to be a good learner, what does that have to do with probability? Well, there is a clue to the answer in the earlier discussion of external calibration and accuracy. There I conjectured that the goal of epistemic rationality was to generate the most accurate, externally calibrated degrees of confidence that we could. The MERF principle is a principle that governs a metacognitive equilibrium process aimed at that goal. Suppose we are successful in reaching that goal—that is, we have fairly accurate, externally calibrated degrees of confidence. To say that they are externally calibrated is to say that whenever we assign confidence of c to a proposition, it is a member of a set of propositions (the set of propositions to which we assign confidence of c) of which the expected relative frequency of truth is c. This defines an empirical probability function. So, if I am right, the goal of epistemic rationality is to make us a reasonably reliable, externally calibrated *empirical* probability generator. To make us good learners in any domain, including logic and epistemology, the requirements of epistemic rationality cannot be probabilist, because they must not impose content restrictions on our beliefs or degrees of confidence that would limit what we could learn—for example, that would prevent us from learning that there are exceptions to the law of excluded middle or that would prevent us from revising our epistemological principles.

Of course, it is a mistake to think that being epistemically rational can itself guarantee success. Epistemic rationality is not a success concept. The example of Descartes' Evil Demon shows that we can be as rational as possible and fail miserably to be reasonably accurate, externally calibrated probability generators. But it is no accident that being epistemically rational is typically a good way to achieve, or approximate our goal.

So I agree that I need a concept of probability to state what the goal of epistemic rationality is. And de Finetti was right, we only need one concept of probability in epistemology. He just got it backward. The only concept we need in epistemology is a concept of empirical probability (e.g., Hoefer 2007).

11 My idealizations

It is important for me to reiterate that I am not criticizing probabilism for incorporating idealizing assumptions into its models of rationality. Almost all explanatory theories involve idealizations, even theories in physics (Cartwright 1983). My own non-probabilist model also depends on idealizing assumptions. In this section I discuss the most important of my idealizing assumptions.

First, some general comments on models.[27] I don't expect any model to explain all of the relevant phenomena. As I see it, models are typically useful in some contexts and not in others or for some purposes and not others. My criticism of probabilist models is not that they don't explain everything, but that there is a large range of cases of metacognitive reasoning that they could not explain and once we have a non-probabilist model that explains those cases, the very same model explains the other cases of metacognitive reasoning, also; so there are no cases of metacognitive reasoning left over for the probabilist model to explain.

In models, simplifying assumptions can make models less realistic, but they can make the logic of the situation more perspicuous. As a general rule, I favor simpler, less realistic models over more complex, more realistic models, if the simpler models illuminate the phenomena of interest. One illustration of this strategy is my use of the special version of the MERF principle in the text, because it is simpler and easier to

[27] For a more thorough and insightful discussion of the role of models in epistemology, see Titelbaum (2013, chaps. 2 and 3).

apply than the general version, which I discuss in the Appendix. I now review some of my other idealizations.

11.1 Point-valued degrees of confidence

For simplicity, in my model, I have assumed point-valued confidence assignments (i.e., an assignment of real numbers) to propositions. For many purposes, as Keynes (1952) has argued, such an assumption is hopelessly unrealistic. Keynes favored interval-valued probabilities rather than point values and this assumption is very common in the literature (e.g., Levi 1980; Kyburg 2003). In some cases, the specification of an interval value alone will not be adequate, it may be necessary to specify a density function for the values in the interval. All of these complications would make my discussion of the examples I am interested in much more complex, with no additional pay-off in understanding. The assumption of point values makes the mathematics trivial so that the philosophical lessons of the examples can be easily grasped. Thus, although point-valued confidence assignments are an extreme idealization, they make it easy to model the change from being almost certain of p (represented by confidence of .99 to P) to being almost certain of −P (represented by confidence of .99 to −P), on discovering that one's confidence in P is entirely due to wish fulfillment.

For cases and issues that require interval-valued confidence assignments, it would be necessary to employ a modified MERF principle that defined *metacognitive disequilibrium* in terms of interval-valued rather than point-valued confidence assignments: a confidence assignment of interval i to P would be in metacognitive disequilibrium if the MCS's estimate of the relevant relative frequency of truth had an interval-value j, and i and j were not equal (or approximately equal). In such a case, equilibrium could be restored by replacing the interval i confidence assignment to P with interval j.

Notice that, no matter what assumption I make about the values of assignments of confidence, whether point values or interval values or a density function over an interval, or something else, they all involve idealizations. There is no way to eliminate idealizations entirely. So the question is not ever whether or not to employ an idealization, but rather what idealization is appropriate for the given explanatory context.

11.2 Identification of reliability-relevant classifications of cognitive processes

In my model, I just assume that the MCS is able to make reliability-relevant classifications of cognitive processes and to distinguish them from other ways of classifying the output of its cognitive processes. I have no theory of reliability-relevance. The term *reliability-relevant* is simply a name for the kinds of classifications that are appropriate for the MCS to employ in its reliability-terminations and an acknowledgment that many classifications are not appropriate. For example, consider Richard's confidence assignment of .99 to P, which he subsequently discovers to be based solely on wish fulfillment. Richard's MCS judges the expected relative frequency of truth of propositions assigned confidence of .99 due solely to wish fulfillment to be .01 and, as a result, reduces his confidence assignment to P to .01. Richard's MCS realizes of course, that it is possible to define a set of propositions (including P) in which the expected frequency of truth is very close to 1.0—for example, a set containing 1000 truths of arithmetic and P—but ignores that fact, because his MCS judges that his confidence assignments to those propositions are the result of very different cognitive processes, with very different relative frequencies of truth.

I must emphasize that the MCS's classification of cognitive processes into kinds is typically very primitive. The MCS may have very little or even a mistaken idea about the details of the relevant processes. So, for example, most people's MCSs make judgments about the reliability of different kinds of memory with practically no information about the kinds of processes involved. The MCS makes certain groupings of confidence assignments (e.g., memories from one's childhood or eyewitness identifications of strangers) based on their seeming to be the result of similar cognitive processes and projects an expected relative frequency on the basis of past errors and general information about the reliability of confidence assignments in the relevant groups.

11.3 The well-behavedness assumption

The two versions of the MERF principle can only be applied if the MCS's reliability determinations are well-behaved, in the sense explained in footnote 8. This seems like an extreme idealization, because in lots of examples from statistics, the relevant empirical probabilities are not well-behaved. For example, Joe is a left-handed Englishman. He can find statistics for the probability at birth that a male resident of England will survive to age 80 and he can find statistics on the probability at birth that a left-handed male resident of Southern California will survive to age 80.[28] Suppose the probabilities are as follows:

prob(live to 80/male resident of England) = .25
prob(live to 80/left-handed male resident of Southern California) = .025

For the probabilities to be well-behaved, Joe would have to have statistics on the life expectancy of left-handed men in England. But suppose there are no relevant data. How should Joe use the data he has to determine a life expectancy for himself? There is no generally accepted solution to this problem.[29]

Because cases like this are very common in statistics, it may seem that my assumption of well-behavedness would greatly limit the applicability of my theory of metacognitive rationality. However, when I reflect on the typical cases of metacognitive reasoning—for example, revising one's degree of confidence in memories of a certain kind or in the results of manual calculations—it seems to me that the MCS typically is able to make an implicit categorization that leads to a fairly precise estimate of the relevant expected relative frequency of truth (e.g., greater than .99, or close to .9 or close to .5). Whether or not these expected relative frequencies of truth should be modeled as point values or as interval values depends on the explanatory context.

[28] In data from Southern California, the life expectancy of left-handed men was 11 years less than the life expectancy of right-handed men (Coren and Halpern 1991).

[29] For a discussion of why it is a problem for all accounts of probability, see Hajek (2007).

Undoubtedly, there will be some cases in which the MCS has statistical evidence that is not well-behaved. Consider, for example, an agent trying to determine the reliability of her eyewitness identification of a robbery suspect who used a gun in a robbery that took place at night. Suppose she has data on the difference in reliability between ordinary eyewitness identifications of a person originally seen in daylight and at night, and also knows that eyewitness identifications in daylight are 10 % less reliable when a weapon is involved. Even though she lacks information about the reliability of eyewitness identifications at night when a weapon is involved, other things being equal, I would expect her MCS to estimate the expected relative frequency of truth of her eyewitness identifications of robbery suspects who used a gun in a nighttime robbery to be approximately equal to 90 % of the reliability of eyewitness identifications at night. This is an example of one of the many ways that the MCS can use statistical information that is not well-behaved to generate expected relative frequencies of truth that are.

However, I do not rule out the possibility that there are cases in which the relevant expected relative frequencies of truth are not well-behaved. In such cases, it may be necessary to employ degrees of confidence with interval values. Because the assumption of well-behavedness generally holds in the cases of interest, I employ it here.

Acknowledgments This paper has benefited from discussions with many people. I received helpful comments on earlier versions from Arthur Fine, Tyler Hildebrand, John Manchak, and Conor Mayo-Wilson, as well as from several anonymous referees. I am grateful to have had the opportunity to work on this paper while in residence as a fellow at the Helen Riaboff Whiteley Center

Appendix

The general MERF principle

To state the general version of the MERF principle I need to introduce two complications to the special MERF principle in the text. The special version of the principle presupposed that, at the meta-level, the agent's MCS has *beliefs* about the relevant factors. This simplified the exposition. The general version of the MERF principle replaces meta-level beliefs with meta-level confidence assignments.

Also, I may have given the impression that there is a division in the cognitive self—that there are two different selves involved in two different levels of cognitive processing. This is only a useful fiction. There is only one self. That single, unified self can evaluate not only the reliability of its ground-level processes; it can evaluate the reliability of its meta-level processes, though of course, it can't evaluate them all at once. The difference in levels is just a heuristic to remind us that whenever the self evaluates its cognitive processes, it steps back and brackets their outputs, rather than just reasserting them.

Here then is the general principle:

> *(General metacognitive expected relative frequency (MERF) Principle) An agent's confidence assignment of c to a proposition P is in disequilibrium if: The agent assigns higher-order confidence to propositions of the following form: There exists*

> *a reliability-relevant category of cognitive processes CP_1 that includes the cognitive processes responsible for the agent's confidence assignment of c to P such that: $ERF(T_r/conf = c, CP_1) = x$;*
> *and the weighted average of these higher-order confidence assignments (the sum of her confidence in each of them weighted by x) is equal to d;*
> *and it is not the case that $d \approx c$;*
> *UNLESS [Narrower Reference Class Exception] the agent assigns higher-order confidence to propositions asserting that there is a reliability-relevant categorization CP_2 of the causal processes responsible for her confidence assignment to P, $CP_2 \leq CP_1$, such that:*
> *$ERF(T_r/conf = c, CP_2) = y$;*
> *and the weighted average of these higher-order confidence assignments (the sum of her confidence in each of them weighted by y) is approximately equal to c.*[30]

This general version of the MERF principle makes it possible to explain how it could be rational for an agent, Van, to be certain about nothing, for it makes it possible for Van's MCS to reduce all of his confidence assignments of 1.0, both ground-level and meta-level, to a value between 0 and 1.0 (and, similarly, to increase all of his confidence assignments of 0 to a value between 0 and 1.0). Here is a simple example meant only to illustrate the main idea: Let Φ be the set of all propositions to which Van assigns confidence of 1.0. I explain how the general MERF principle could lead him to adopt a confidence assignment in which every member of Φ is assigned confidence of .998, with the result that no proposition is assigned confidence of 1.0 (or 0). For simplicity, I assume that there is no relevant narrower subclass Φ for which Van's MCS projects a different expected relative frequency of truth than for Φ. In the cases of interest, Van will not assign confidence of 1.0 to the proposition that the expected relative frequency of truth of the members of Φ is .998. Van will divide his confidence among various alternative values for that expected relative frequency of truth. Here is a simple example: Van assigns confidence of .5 to each of two possibilities: that the relevant relative frequency of truth is .999 and that the relevant relative frequency of truth is .997. His confidence assignment of 1.0 to a proposition P (in Φ) will be in disequilibrium when, as in this case, the weighted average of his various estimates of the expected relative frequency of truth in the members of Φ (in this case, .998) is not equal to his confidence in the individual members of Φ (in this case, 1.0). In the simplest case, the general MERF principle will require him to reduce his confidence assignment to each of the members of Φ to .998.

A proof that probabilist epistemologies are immodest

To show that all probabilist epistemologies imply that we are rationally required to assign confidence of 1.0 to at least some substantive epistemological claims, I carry

[30] The general MERF Principle is a refinement of an earlier proposal (Talbott 1990, Preface 31–35). As before, I assume that the agent's confidence assignments to the relevant expected relative frequencies are well-behaved, in the sense explained in footnote 8, when suitably generalized by replacing references to *beliefs* about the relevant expected relative frequencies with references to *confidence assignments* to the relevant expected relative frequencies.

out the argument for strict classical probabilism. It is easily modified to apply to non-strict classical probabilism, and even to non-classical probabilism, simply by making suitable substitutions for the variable 'L,' because all of these views require certainty in some propositions, and those requirements support implications to the conclusion that some substantive epistemological claims must be certain. Here is the argument for strict, classical probabilism:

> Let *RC(P) = x* be the relation: Rationality requires assigning confidence of x to proposition P. The central principle of any form of probabilism is:
> (1) [prob (P) = x] iff [RC(P) = x].
> Consider L, a truth of classical logic. Strict classical probabilism requires:
> (2) prob(L) = 1.
> From (1):
> (3) [prob(L) = 1] iff [RC(L) = 1].
> From (2) and (3):
> (4) RC(L) = 1 [Rationality requires assigning confidence of 1.0 to L.]
> What is the probability of (4)? To answer this question, we need to determine the probability of (2)—that is, the probability of a probability, which is a higher-order probability. In a formalism rich enough to coherently model higher-order probabilities (e.g., Skyrms 1980), from (2) it follows that:
> (5) prob[prob(L) = 1] =1.
> We also need one more probability theorem:
> (6) [P iff Q] → [prob(P) = prob (Q)].
> Then from (3), (5), and (6):
> (7) prob[RC(L) = 1] = 1.
> And from (1) and (7):
> (8) RC([RC(L) = 1] = 1.

So strict classical probabilism requires an assignment of confidence of 1.0 to it's own non-trivial epistemological claim [RC(L) = 1]—that is, it requires a confidence assignment of 1.0 to the claim that we are rationally required to assign confidence of 1.0 to L. We can continue the construction to generate a potentially infinite list of non-trivial claims of strict classical probabilist epistemology to which strict classical epistemology requires us to assign confidence of 1.0. This shows that strict classical probabilism is an *immodest* epistemology.[31] Parallel arguments show that any non-strict classical probabilist epistemology, such as Garber (1983), and even any non-classical probabilist epistemology must be immodest. So all probabilist theories, which include all Bayesian theories, are immodest. They all require rational degrees of confidence of 1.0 in at least some of their own substantive (i.e., non-trivial) epistemological claims.

[31] Sobel (1987) uses Dutch Book arguments to defend this kind of immodesty as a virtue of probabilist theories. I discussed Dutch Book arguments above. For another kind of Bayesian immodesty, see Belot (2013).

A proof that departures from MERF-Defined equilibria increase expected inaccuracy

In the text, I assert that when an agent S's confidence assignment satisfies the special or general MERF Principle and S's MCS has opinions about the relevant relative frequencies of truth, changes to S's confidence assignment increase its expected inaccuracy (and thus decrease its expected accuracy). Here I prove this result for the most commonly used measure of inaccuracy, the Brier score, according to which the inaccuracy of a confidence assignment of x to P equals the square of its distance from the truth value of P (i.e., 1.0 if p is true; 0 if p is not true).

The key step in the proof is to define the probabilities to be used in the definition of expected inaccuracy. Joyce correctly points out that expected inaccuracy cannot be consistently defined using non-probabilist confidence assignments in the role of probabilities (1998, pp. 589–590). I do not do so. The probabilities I use are the probabilities defined by the MCS's expected relative frequency of truth for the narrowest relevant reference class that includes the proposition of interest. The MCS must have opinions about those relative frequencies for these probabilities to exist. So there can be no determination of expected inaccuracy without them.

I use *prob* to refer to the MCS's relevant estimates of the expected relative frequencies of truth. Then the expected inaccuracy (EI) of an agent S's confidence assignment of z to proposition P can be defined as the weighted sum of its inaccuracy if P is true ($[1-z]^2$), weighted by the probability that P is true [prob(P)] and its inaccuracy if P is not true (z^2), weighted by the probability that P is not true (prob(−P) = [1−prob(p)]).[32]

In the case in which S's confidence assignment of x to P is in equilibrium, prob(P) = conf(P) = x. Therefore, the expected inaccuracy of the confidence assignment of x to P is:

(1) $\mathrm{EI}(\mathrm{conf}(p) = x) = x[1-x]^2 + (1-x)x^2 = x - x^2$

I compare (1) with what the expected inaccuracy of S's confidence assignment to P would be if S's confidence in P were increased from x to [x+y] (where $y > 0$ and $0 \leq [x + y] \leq 1$). (The proof of the case in which S's confidence assignment to p is decreased is exactly parallel.) Intuitively, increasing S's confidence assignment to P will decrease the inaccuracy of the assignment if P is true and increase the inaccuracy of the assignment if P is not true. The expected inaccuracy of S's confidence assignment of [x+y] to P is again a weighted sum of two components: the inaccuracy of the assignment of [x+y] to P if P is true ($[1 - [x + y]]^2$), weighted by the probability that P is true (in this case, x), and the inaccuracy of the assignment of [x + y] to P if P is not true ($[x + y]^2$), weighted by the probability that P is not true (1−x). Thus:

(2) $\mathrm{EI}(\mathrm{conf}(p) = [x+y]) = x(1 - [x + y])^2 + (1 - x)[x + y]^2 = x - x^2 + y^2$

The expected inaccuracy of the confidence assignment of [x + y] to P is greater than the expected inaccuracy of the confidence assignment of x to P by the amount y^2.

[32] The only novel part of the proof is to interpret the expected relative frequencies of truth as probabilities by reference to which expected inaccuracy can be defined. The rest of the proof simply follows de Finetti's [1940] that the Brier Score is a proper scoring rule.

So the confidence assignment of x to P minimizes expected inaccuracy, and the farther S's confidence to P departs from x (i.e., |y|), the greater its expected inaccuracy.

References

Belot, G. (2013). Bayesian orgulity. *Philosophy of Science*, *80*, 483–503.

BonJour, L. (1998). *In defense of pure reason*. Cambridge: Cambridge University Press.

Cartwright, N. (1983). *How the laws of physics lie*. Oxford: Clarendon Press.

Christensen, D. (2007). Does Murphy's law apply in epistemology? Self-doubt and rational ideals. *Oxford Studies in Epistemology*, *2*, 3–31.

Christensen, D. (2010a). Higher-order evidence. *Philosophy and Phenomenological Research*, *81*, 185–215.

Christensen, D. (2010b). Rational reflection. *Philosophical Perspectives*, *24*, 121–140.

Christensen, D. (2011). Disagreement, question-begging, and epistemic self-criticism. *Philosophers' Imprint*, *11*, 1–21.

Coren, S., & Halpern, D. F. (1991). Left-handedness: A marker for decreased survival fitness. *Psychological Bulletin*, *109*, 90–106.

Crick, F. (1988). *What mad pursuit*. New York: Basic Books.

de Finetti, B. [1937] (1980). La Prevision: Ses Lois Logiques, Ses Sources Subjectives. (*Annales de l'Institut Henri Poincare*, *7*, 1–68). Translated into English and reprinted in Kyburg and Smokler, *Studies in Subjective Probability*. Huntington: Krieger.

de Finetti, B. [1940] (2008). Decisions and proper scoring rules. In Alberto, M. (Ed.), Hykel Hosni, tr., *Philosophical Lectures on Probability* (pp. 15–26). Dordrecht: Springer.

Feldman, R. (1985). Reliability and justification. *Monist*, *68*, 159–174.

Feldman, R. (2005). Respecting the evidence. *Philosophical Perspectives*, *19*, 95–119.

Field, H. (1996). The A prioricity of logic. *Proceedings of the Aristotelian Society*, *96*, 359–379.

Field, H. (2008). *Saving truth from paradox*. Oxford: Oxford University Press.

Garber, D. (1983). Old evidence and logical omniscience in bayesian confirmation theory. In J. Earman (Ed.), *Testing scientific theories, midwest studies in the philosophy of science 10* (pp. 99–131). Minneapolis: University of Minnesota Press.

Glymour, C. (1980). *Theory and evidence*. Princeton: Princeton University Press.

Hajek, A. (2007). The reference class problem is your problem too. *Synthèse*, *156*, 563–585.

Hoefer, C. (2007). The third way on probability: A Sceptic's guide to chance. *Mind*, *116*, 549–596.

Horowitz, S. (2014). Epistemic akrasia. *Noûs*, *48*, 718–744. doi:10.1111/nous.12026.

Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach*. Chicago: Open Court.

Jeffrey, R. (1983). Bayesianism with a human face. In J. Earman (Ed.), *Testing scientific theories, minnesota studies in the philosophy of science* (Vol. 10, pp. 133–156). Minneapolis: University of Minnesota Press.

Jeffrey, R. (1986). Probabilism and induction. *Topoi*, *5*, 51–58.

Jeffrey, R. (1992). *Probability and the art of judgment*. Cambridge: Cambridge University Press.

Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science*, *68*, 575–603.

Kelly, T. (2010). Peer disagreement and higher-order evidence. In R. Feldman & T. A. Warfield (Eds.), *Disagreement* (pp. 111–174). Oxford: Oxford University Press.

Keynes, J. M. (1952). *A treatise on probability*. London: Macmillan.

Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy*, *72*, 690–716.

Kyburg, H. E, Jr. (2003). Are there degrees of belief? *Journal of Applied Logic*, *1*, 139–149.

Laudan, L. (1981). A confutation of convergent realism. *Philosophy of Science*, *48*, 19–48.

Levi, I. (1980). *The enterprise of knowledge*. Cambridge: MIT Press.

Levi, I. (1991). *The fixation of belief and its undoing*. Cambridge: Cambridge University Press.

Lewis, D. (1980). A subjectivist's guide to chance. In R. C. Jeffrey (Ed.), *Studies in inductive logic and probability* (Vol. 2, pp. 263–293). Berkeley: University of California Press.

Millgram, E. (2009). *Hard truths*. Chichester: Wiley-Blackwell.

Paris, J. B. (2001). A note on Dutch book method. *Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications* (pp. 301–306). Ithaca: Shaker.

Pollock, J. L. (1984). Reliability and justified belief. *Canadian Journal of Philosophy*, *14*, 103–114.

Priest, G. (2002). Paraconsistent logic. In D. Gabbay & F. Guenthner (Eds.), *Handbook of philosophical logic* (2nd ed., Vol. 6, pp. 287–393). Dordrecht: Kluwer Academic Publishers.

Quine, W. V. O. (1961). Two dogmas of empiricism. In W. V. O. Quine (Ed.), *From a logical point of view* (pp. 20–46). New York: Harper & Row.

Seidenfeld, T., Schervish, M. J., & Kadane, J. B. (2012). What kind of uncertainty is that? Using personal probability for expressing one's thinking about logical and mathematical propositions. *Journal of Philosophy*, *109*(2012), 516–533.

Shimony, A. (1988). An adamite derivation of the principles of the calculus of probability. In J. H. Fetzer (Ed.), *Probability and causality* (pp. 79–89). Dordrecht: Kluwer.

Skyrms, B. (1980). Higher order degrees of belief. In D. H. Mellor (Ed.), *Prospects for pragmatism* (pp. 109–137). Cambridge: Cambridge University Press.

Skyrms, B. (1990). *The dynamics of rational deliberation*. Cambridge: Harvard University Press.

Sobel, J. H. (1987). Self-doubts and Dutch strategies. *Australasian Journal of Philosophy*, *65*, 56–81.

Talbott, W. J. [1990]. (2015). *The reliability of the cognitive mechanism: A mechanist account of empirical justification*. New York: Routledge.

Tarski, A. (1944). The semantic conception of truth. *Philosophy and Phenomenological Research*, *4*, 341–376.

Titelbaum, M. (2013). *Quitting certainties: A Bayesian framework modeling degrees of belief*. Oxford: Oxford University Press.

Titelbaum, M. (2015). Rationality's fixed point: (or: In defense of right reason). In T. S. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology* (Vol. 5, pp. 253–294). Oxford: Oxford University Press.

Van Fraassen, B. C. (1983). Calibration: A frequency justification for personal probability. In R. Cohen & L. Laudan (Eds.), *Physics, philosophy, and psychoanalysis* (pp. 295–319). Dordrecht: D. Reidel.

Williams, J. R. G. (2012). Gradational accuracy and non-classical semantics. *Review of Symbolic Logic*, *5*, 513–537.