Being Rational and Being Wrong

Kevin Dorst
kevindorst@pitt.edu

Draft—comments welcome!

September 2020

Abstract

Do people tend to be overconfident in their opinions? Many think so. They've run studies to test whether people are *calibrated*: whether their confidence in their opinions matches the proportion of those opinions that are true. Under certain conditions, people are systematically "over-calibrated"—for example, of the opinions they're 80% confident in, only 60% are true. From this observed over-calibration, it's inferred that people are irrationally *overconfident*. My question: When—and why—is this inference warranted? Answering this question requires articulating a general connection between being rational and being right—something extant studies have not done. I show how to do so using the notion of *deference*. This provides a theoretical foundation to calibration research, but also reveals a flaw: the connection between being rational and being right is much weaker than is commonly assumed; as a result, rational people can often be expected to be miscalibrated. Thus we can't test whether people are overconfident by simply testing whether they are over-calibrated; instead, we must first predict the expected *rational deviations* from calibration, and then compare those predictions to people's performance. I show how in principle this can be done—and that doing so has the potential to overturn the standard interpretation of robust empirical effects. In short: rational people can be expected to be wrong more often than you might think.

1 The Question

Pencils ready! For each pair, circle the city that you think has a larger population (in the city proper), and then rate how confident you are in that guess on a $50-100\%$ scale:

1) Denver or Phoenix? Confidence: ____%

2) San Jose or Seattle? Confidence: ____%

3) Indianapolis or Columbus? Confidence: ____%

If you're like most people, this test will reveal two things. First, it's likely that only one or two of your answers is correct. Second—and perhaps more worryingly—it's likely that your confidence in your answers does not match this probability of being correct. Among 200 test-takers, the average confidence people had in their answers was 75%, while the proportion of correct answers to hard questions like this was only 45%.[1]

That rather striking result—the so-called "overconfidence effect"—is common: on a variety of tests, people's average confidence in their answers exceeds the proportion that are correct.[2] Many have concluded from this result that people are often overconfident in their opinions—i.e. more confident than it is rational for them to be, given their evidence.[3] Many have used these (and related) results to paint unflattering pictures of the human mind as prone to pervasive irrationality and bias.[4] And many others have invoked overconfidence in particular to explain a variety of societal ills—from market crashes, to political polarization, to wars.[5] Daniel Kahneman summed it up bluntly: 'What would I eliminate if I had a magic wand? Overconfidence' (Shariatmadari 2015).

Okay. But how—exactly—did we reach this conclusion of pervasive overconfidence?

The evidence comes in the form of calibration studies like the one you just took. We ask people a variety of questions, have them report their confidence in their answers, and then graph that confidence against the proportion of answers that are true.[6] Say that a person is *calibrated* (at x) if exactly x% of the claims that they are x% confident in are true. They are *over*-calibrated (at x) if fewer than x% of such claims are true.[7] And they are *under*-calibrated (at x) if more than x% of such claims are true. Focusing on binary-question ("2-alternative-forced-choice") formats—wherein people are asked to choose between two answers, and so are always at least 50% confident in their answer—schematic graphs of these different **calibration curves** are given on the left of Figure 1. Meanwhile, the right side of Figure 1 plots the results of my study (see §5.2 for details), replicating a well-known result that (on certain types of questions) people tend to be

[1] Answers: Phoenix, San Jose, Columbus. "Hard questions" means those with hit rates below 75%; see §5 for study details.

[2] For summaries, see Lichtenstein et al. (1982), Harvey (1997), Hoffrage (2004), Glaser and Weber (2010), and Moore et al. (2015).

[3] Lichtenstein et al. (1982); Dunning et al. (1990); Vallone et al. (1990); Griffin and Tversky (1992); Kahneman and Tversky (1996); Budescu et al. (1997); Brenner (2000); Koehler et al. (2002); Brenner et al. (2005); Glaser and Weber (2010); Merkle and Weber (2011); Brenner et al. (2012); Moore et al. (2015); Ehrlinger et al. (2016); Magnus and Peresetsky (2018).

[4] Plous (1993); Fine (2005); Ariely (2008); Hastie and Dawes (2009); Myers (2010); Kahneman (2011b); Thaler (2015); Lewis (2016); Tetlock and Gardner (2016).

[5] E.g. Howard (1984); Odean (1999); Glaser and Weber (2007); Johnson (2009); Myers (2010, 377); Johnson and Fowler (2011); Kahneman (2011a); Ortoleva and Snowberg (2015); van Prooijen and Krouwel (2019).

[6] I'll focus on this type of calibration study—but see §6 for sketches of how the lessons may apply to both placement- (Kruger and Dunning 1999) and interval-estimation (Moore et al. 2015) methods.

[7] "Over"-calibrated because their confidence in those opinions needs to be lower to be calibrated. In the graphs below, imagine the person controlling a left-right slider for their confidence; over-calibration is putting it too far to the right; under-calibration is putting it too far to the left.

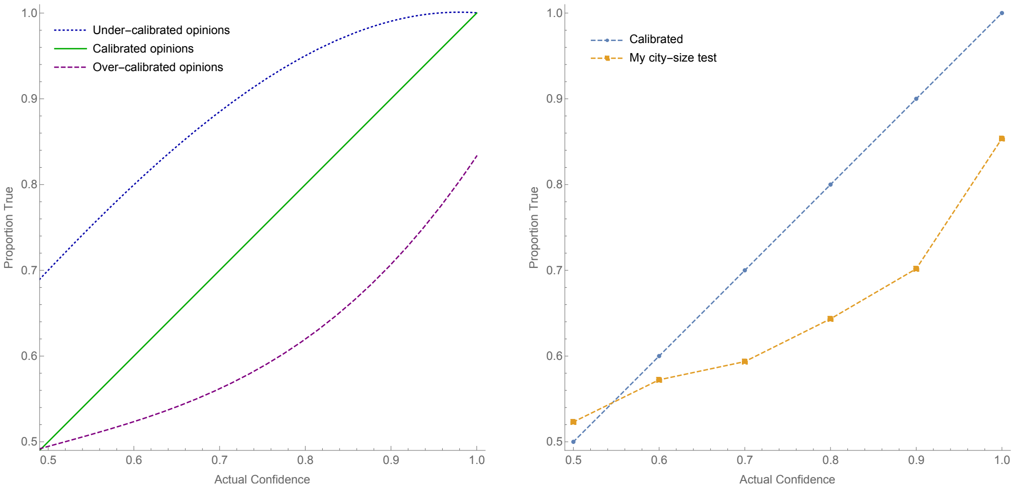

Figure 1: **Left:** Schematic calibration curves. **Right:** "Overconfidence effect" (i.e. over-calibration) in my study. (See §5.2 for study details.)

substantially over-calibrated (Lichtenstein et al. 1982).

That's the evidence—namely, that people are often over-calibrated. How do we get from there to the conclusion—namely, that people are often overconfident? Well, it's natural to think that if people's confidence is rationally placed, their opinions will be right about as often as they expect them to be. Conversely, if people are quite confident in their opinions and yet many (or most!) of those opinions are wrong, it's natural to infer that they are *too* confident—*over*confident.

This is natural. But it is also a substantive inference: it moves from an empirical observation—'you are (mis)calibrated'—to a normative conclusion—'you are (ir)rational.' Call it the **rational-to-right inference** since, stated bluntly, it relies on a general connection between your opinions being rational and your opinions being right.

The Questions: What *is* the connection between being rational and being right? More specifically: When is the rational-to-right inference warranted? When is it not? And what does that tell us about how to interpret the results of calibration studies?

The Plan: I'll first say what sort of connection the rational-to-right inference assumes, and explain why the existing literature has failed to articulate it (§2). I'll then go on to use the notion of *deference* to articulate this connection, and show how it vindicates the rational-to-right inference in certain simple cases (§3). However, it turns out that even the strongest (most contentious) version of this connection will break in predictable ways—meaning that often *mis*calibration is evidence for *rationality* (§4). I conclude by arguing that this result provides both a foundation for and a refinement to the standard calibration-study methodology: in testing whether people are rational, the null hypothesis should not be that they will be calibrated; rather, we must first predict the rational *deviations* from calibration on our test, and then compare people's performance to those predictions. I show how in principle this can be done, and that doing so has the potential to overturn the interpretation of robust empirical effects (§§5–6).

The Upshots: If all this is correct, it shows that certain philosophical and psychological literatures are much more intimately connected than has been realized. Contemporary philosophical debates about the formulation and tenability of deference principles have a direct and substantive bearing on the methodology and interpretation of empirical studies of confidence. Conversely, the methods developed by these studies show how we can make precise predictions about the relationship between rationality and truth in a variety of environments—and that being rational is not nearly the guide to being right that you might think. Regardless of whether you accept these particular conclusions, I hope to convince you that there are rich connections here—and thus that philosophers, psychologists, and behavioral economists can productively work together more closely in tackling the question of human (ir)rationality.

2 The Problem

There is a problem here. The rational-to-right inference involves three quantities:

(1) A person's *actual* degrees of confidence in some claims.
(2) The proportion of those claims that are true.
(3) The degrees of confidence it would be *rational* for them to have in those claims.

The only quantities that are observed are (1) and (2). The rational-to-right inference uses these to infer something about (3): from the observation that (1) is higher than (2), it is inferred that (1) is higher than (3). Clearly this makes sense only if rational confidence—(3)—can be expected to align with proportion true—(2).

The point can be made graphically. What would it mean to say that people tend to be overconfident (in a given domain[8])? I'll take it to mean that they are (on average) *more extreme* in their opinions in that domain than they would be if they were rational. If we plot actual degrees of confidence against rational degrees of confidence (on $50 - 100\%$ scale), people tend to be rational if (averaging across opinions) rational confidence matches actual confidence—the curve is diagonal; they tend to be overconfident if rational confidence is less extreme than actual confidence—the curve is tilted. (See the left side of Figure 2.) That's the overconfidence hypothesis. What is the evidence offered in its favor? It's that in a variety of settings, people are over-*calibrated*: if we plot actual degree of confidence against *proportion true*, the curve is tilted—see the right side of Figure 2.

[8] The "in a given domain" rider is important, as patterns of miscalibration vary widely across different sets of questions (see Koehler et al. 2002; Brenner et al. 2005). We'll introduce refinements to the empirical story in §4.1 and onwards—for now, I'll focus on tests for which the "overconfidence effect" is observed.

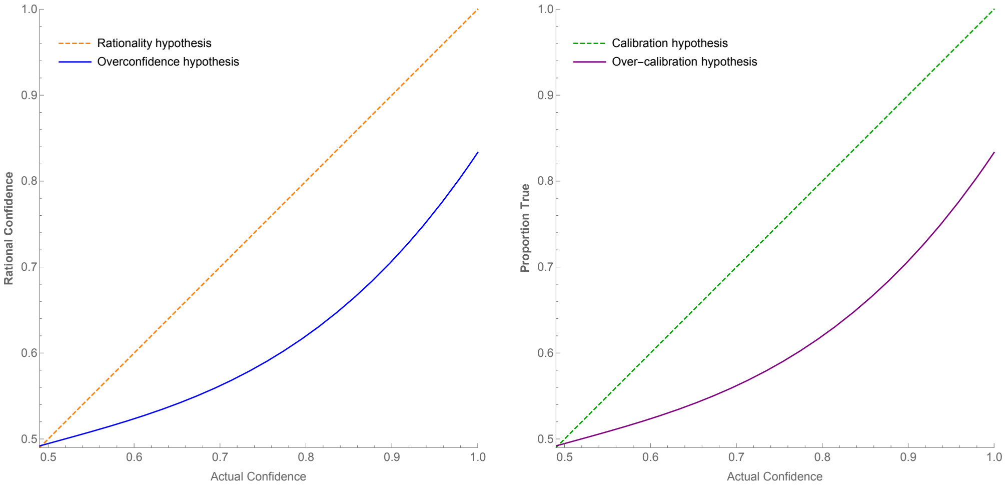

Figure 2: **Left:** Rationality vs. Overconfidence hypotheses. **Right:** Calibration vs. Over-calibration hypothesis.

Simple point: although the graphs look the same, the axes are different. It follows that the rational-to-right inference is warranted when and only when you should expect the two axes to align—i.e. when you should expect a rational person's judgment to be calibrated on the given test. More specifically, the inference works when and only when the following hold. Take all the claims that the person *should* be 50% confident in—you should expect roughly half of them to be true; take all the claims that the person *should* be 60% confident in—you should expect roughly 6 out of every 10 of them to be true; and so on.

There's a point here worth emphasizing. To say that someone is overconfident in a set of opinions $q_1, ..., q_n$ is to say that they are, on average, more confident than they *should* be—that is, that there is some number c that represents their average confidence, some other number r that represents the average confidence it would be rational for them to have, and that $c > r$. What this means is that calibration studies—and the rational-to-right inference they are based upon—presuppose that there are rational degrees of confidence (r_i) that people ought to have in the claims they evaluate (q_i), which may differ from their reported degrees of confidence (c_i).

Why am I banging on this drum? Because *I know of no study that explicitly represents the rational degrees of confidence r_i as variables to be investigated.* None of the studies cited in this paper do so.[9] As a result, none of these studies state the assumptions needed about rational confidence to derive the result that we should expect the

[9] Including those cited in footnotes 2, 3, and 5. Some studies model notions of probability distinct from subjects' reported confidence and observed frequencies—such as objective probabilities, true subjective confidence (as distinct from reported confidence), or the subjective confidence of differing agents (Gigerenzer et al. 1991; Erev et al. 1994; Juslin et al. 1997, 1999, 2000; Moore and Healy 2008). None of these quantities are treated as rational confidence needs to be—see below for more discussion.

rational opinions r_i to be calibrated in their study.[10] In other words: I know of no study that states what assumptions it is making such that we should expect the two y-axes in Figure 2 to align. Yet *over-calibration is evidence for overconfidence only if we should expect them these axes to align*: observing that people's judgments are miscalibrated provides no evidence that they are irrational unless we have reason to think that the *rational* degrees of confidence would be calibrated. That is the problem.[11]

But how serious is this problem? Can we safely assume that—at least if the study is properly set up—rational confidence will on average be calibrated?

No: there is no necessary connection between being rational and being right at any level of statistical generality.

This is easy to see in extreme cases. *Case 1:* Take the philosopher's favorite rational brain-in-a-vat, Rajat. Rajat rationally uses all the information he gets. His information is a lot like yours or mine. As a result he's sure that he has hands, confident he's healthy, and suspects he'll soon grab lunch. But though rational, Rajat is wrong on all these fronts (and many others)—for, unbeknownst to him, he's a brain-in-a-vat being deceived by a mad scientist into thinking he's living a normal life. If we ran a calibration study on Rajat, he would be systematically over-calibrated—most of the things he's confident in are false. Yet, by stipulation, we know Rajat is perfectly rational.

More mundane cases make the point as well. *Case 2:* Meet Georgie. She's quite confident—and quite wrong—in most of her geographical opinions. When she took a city-population test, her average confidence was 90% in her guesses, but the proportion she answers correctly was 50%. Does this provide evidence that she was irrationally overconfident? Not if we know that her geography teacher gave her an outdated textbook on city-sizes to memorize, for then we should chalk up her mistakes to bad information rather than irrationality.

Obviously we can imagine scenarios in which the entire population of test-takers are the same position—we'd expect *every* student in Georgie's geography class to be rationally highly confident in their opinions, and yet also systematically wrong.

Likewise, it's easy to construct cases in which we know the subject's have high-quality evidence, and yet the rational-to-right inference fails. *Case 3:* I have a coin in my pocket that's 60% biased toward heads; I'm about to toss it 100 times. How

[10] Some explicitly derive this result for a given Bayesian agent (Brenner et al. 2005; Moore and Healy 2008; Merkle and Weber 2011)—but to do so they all implicitly assume that the Bayesian's prior beliefs match the objective frequencies on the test. As we'll see, this cannot in general be assumed.

[11] Lest you wonder if this suggests that psychologists are not interested in rationality, and instead are interested purely in the descriptive phenomenon of over-calibration, rest assured that the normative interpretation of these studies is clear. They are peppered with normative assessments of people's confidence: e.g. 'irrational" (Hoffrage 2004, 245; Magnus and Peresetsky 2018, 2), "unjustified" (Dunning et al. 1990, 579; Vallone et al. 1990, 588), "unreasonable" (Merkle and Weber 2011, 264), "biased" (Koehler et al. 2002, 686; Glaser and Weber 2010, 249; Moore et al. 2015, 182), etc. Kahneman and Tversky put it bluntly: "Our disagreement [with Gigerenzer (1991)] is normative, not descriptive. We believe that subjective probability judgments should be calibrated, whereas Gigerenzer appears unwilling to apply normative criteria to such judgments" (Kahneman and Tversky 1996, 589).

confident are you, of each toss, that it'll land heads on that toss? Write that number down—I'll look at it in a second. First to toss the coin (...done). Turns out it landed heads only 30 times. Now to compare that to your confidence.... Hm, 60%? You were 60% confident that each toss would land heads, but only 30% of those claims were true. Have I gained evidence that you are overconfident? Obviously not—your 60% confidence was perfectly rational, yet sometimes rational opinions turns out to be mistaken.

Similarly, sometimes you can *know* that your rational opinions will be systematically mistaken. *Case 4:* I have an urn of mis-printed coins—60 of them are double-headed, and the remaining 40 are double-tailed. I'm about to pull a single coin from the urn and toss it 100 times. How confident are you, of each toss, that the coin I draw will land heads on that toss? 60% I take it. Yet you know that either I'll draw a double-headed or a double-tailed coin. If the former, all the tosses will land heads—100% of the things that you're 60% confident in will be true. And if the latter, then none of them will land heads—0% of the things that you're 60% confident in will be true. So you know that, either way, the rational opinions will be badly miscalibrated.

Finally: in almost any conceivable scenario, a rational person will know that *certain classes of their opinions* will be systematically miscalibrated. *Case 5:* Suppose you're about to take a test drawn randomly from a representative set of your knowledge about geography. Suppose you know that the sources you've studied diligently and rationally are generally accurate. Nevertheless, your rational opinions aren't perfect—sometimes you'll be wrong. Consider the set of guesses $\mathcal{W}$ you'll be wrong about, and the set $\mathcal{R}$ you'll be right about. You won't know what these sets are until you finish the test and the answers are revealed. But you *know* you'll be miscalibrated on them—0% of the claims in $\mathcal{W}$ will be true, but your average confidence in them will be higher than that; and 100% of the claims in $\mathcal{R}$ will be true, but you average confidence in them will be lower than that. More generally, we should *always* expect that people will be over-calibrated on sets of opinions like "the set of answers they tend to get wrong" and under-calibrated on sets lke "the set of answers they tend to get right."[12]

Upshot: it is easy to imagine scenarios in which perfectly rational people are systematically miscalibrated.[13] Thus when we run the rational-to-right inference—inferring

[12] Cf. the discussion of "linear dependency" in Juslin et al. (2000) and elsewhere; we'll come back to this point as it relates to the "hard-easy effect" in §5.

[13] For those familiar with certain bits of theory, a clarification may be helpful here. Any Bayesian will expect any particular set of their own opinions to be calibrated (see below). But we are not them, and we know things that they do not. Therefore there is no theorem that *we* should expect them to be calibrated. Often we should not. (Why must *they* expect to be calibrated? Because a Bayesian's estimate of the proportion of truths amongst some particular set of claims will equal their average degree of confidence in them. Letting C be any probability function, $\mathbb{E}[X]$ be its expectation of a variable X ($\mathbb{E}[X] := \sum_t C(X = t) \cdot t$), and $I(q_i)$ be the indicator of q_i (1 if q_i is true, 0 if not), we have: $\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} I(q_i)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[I(q_i)] = \frac{1}{n}\sum_{i=1}^{n} C(q_i)$. Thus our subject's estimate of the proportion of truths amongst the claims they are 80% confident in must be 80%. Moreover, so long as they treat the claims (relatively) independently, they will (by the weak law of large numbers) be confident that roughly 80% of those claims are true.)

from miscalibration to irrationality—we are somehow discounting concerns from scenarios like this. The question is what justifies us in doing so.

To be clear: I am not claiming that these toy cases shed doubt on the rational-to-right inference in practice (nor that they will be of any surprise to the researchers conducting calibration studies!). What I'm claiming is that these cases make salient a conceptual question. As we've seen, whether we should expect the rational opinions to be calibrated on a given set of questions depends completely on the evidence that the test-taker has and how that set was determined. Surely, in some sense, we should expect that the opinions their evidence warrants will tend to be right—that's the point of evidence, after all—and thus that rational degrees of confidence will *tend* to be calibrated. The question is: *When, why, and in what sense should we expect this?*

The answer to this question is not obvious. It requires formulating a systematic connection between being rational and being right. As discussed above, none of the calibration studies I know of have articulated such a connection, for none of them have represented the rational degrees of confidence as a variable to make assumptions and predictions about in their test. That is what I'm going to do. In §3 I'll articulate a general, probabilistic connection between being rational and being right—and on what this connection depends. This explains why the rational-to-right inference works in certain simple cases. But, as we'll see in §4, it also reveals that it'll fail in systematic ways whenever we have information that the test-takers don't—as we always will. §5 will turn to saying what this implies about the proper methodology of calibration studies.

But before moving on, I should say more about how this project relates to an array of theoretical points that have been made in the calibration literature.[14] "Ecological" approaches have made the point that subjects may well have misleading information about our test, and therefore that we must try to control for this by choosing representative questions from a natural domain (Gigerenzer 1991; Gigerenzer et al. 1991; Juslin 1994; Juslin et al. 2000; Hoffrage 2004). "Error-model" approaches have made the point that even if questions are chosen randomly from a natural domain, there will be stochastic errors ("noise") in both the selection of items and in the subject's reporting of their confidence that can naturally lead to them being miscalibrated on a given test—even if their true opinions are calibrated overall (Erev et al. 1994; Pfeifer 1994; Juslin et al. 1997, 1999, 2000). Similar points have been made using information asymmetries between subjects (Moore and Healy 2008; Jansen et al. 2018). Based on such considerations,

[14] What about precedents in the *philosophical* literature? To my knowledge no philosophers have directly addressed the rational-to-right inference, instead focusing on different questions about the epistemic significance of calibration. Some ask whether calibration can objectively vindicate a set of opinions (van Fraassen 1983; Dawid 1983; Seidenfeld 1985; Joyce 1998; Dunn 2015; Pettigrew 2016a); others ask whether a Bayesian agent's beliefs about their own long-run calibration are problematic (Dawid 1982; Belot 2013a,b; Elga 2016); and others ask how our your expectations about your (or your peer's) calibration should affect your confidence in your answers (Roush 2009, 2016, 2017; White 2009b; Christensen 2010a, 2016; Lam 2011, 2013; Sliwa and Horowitz 2015; Schoenfield 2015, 2016a; Isaacs 2019).

these researchers have built models of how people may form their degrees of confidence in an apparently rational way, and yet nonetheless we might expect to see the sorts of miscalibration that we in fact observe.

I agree with these conceptual points, and some of my modeling choices in §5 are inspired by them. But the point I'm making is a broader one.

These researchers have proposed particular, rational-seeming mechanisms for forming opinions[15], and shown that they can lead to miscalibration. What I'm going to show is that *no matter the mechanism* used to form beliefs, rational opinions should be expected to be miscalibrated in systematic ways. Demonstrating this becomes possible once we explicitly represent the rational opinions as variables to be investigated. Interestingly, these rational deviations from calibration turn out to be broadly consistent with some of the core empirical trends (§5.2). But more importantly, they show that *we've been using the wrong yardstick*. In assessing whether people are overconfident, we should never simply compare their calibration curves to the diagonal calibrated line—rather, we must compare them to the predicted rational *deviations* from the calibrated line. I'll show how we can in principle predict these rational deviations from calibration without making any assumptions about mechanism.

If this is right, the sorts of simulations for predicting miscalibration pioneered by Erev et al. (1994); Pfeifer (1994), and Juslin et al. (1997)—and which I will use in §5—are not the special purview of those trying to explain the empirical data with a rational model of confidence. Rather, they are a necessary precondition for figuring out what the null hypothesis should be when we aim to assess whether people's calibration curves provide evidence for overconfidence.

3 The Insight

First things first, we need to delineate the connection between being rational and being right. When—and why—should we expect the rational degrees of confidence for a given person to be calibrated?

When you learn about the results of a calibration study, you get a *lot* of evidence: how (mis)calibrated many subjects were across many levels of confidence; what sorts of test items were used, and how they were selected and presented; etc. All this evidence makes things complicated.

Let's start by making things simple. Suppose you get very limited evidence. A single subject—Calvin—was given a calibration test; the questions were selected to be random and unrelated. Consider all the claims that Calvin was 80% confident in—call those his **80%-opinions**. All you're told is which proportion of them were true.

[15] Which, in turn, have been criticized on a variety of grounds (Kahneman and Tversky 1996; Budescu et al. 1997; Brenner 2000; Koehler et al. 2002; Brenner et al. 2005; Merkle and Weber 2011).

I claim that in this simple scenario, the rational-to-right inference is warranted. If you learn that (roughly) 80% of Calvin's 80%-opinions were true, you get strong evidence that those opinions were rational; if you learn that far fewer (or far more) than 80% of these opinions are true—say, 60% (or 95%)—you get strong evidence that he was overconfident (underconfident). This, I claim, is the insight behind calibration studies. Why is it correct?

Begin with a parable. Long ago, Magic Mary possessed a variety of magic coins—some were biased to come up heads almost every time; others to come up heads 90% of the time; others 80%, and so on. The magic coins had a variety of special markings on them—on some, George Washington has a large nose and small ears; on others, he has a thin neck and bushy eyebrows; etc. In principle, if one knew how to decipher the markings, one could tell what the bias of the coin was just by looking at it.

Mary tossed the coins many, many times. She kept fastidious records: for each toss she wrote the details of the coin's markings on one side of a stone tablet, and the outcome of the toss (heads or tails) on the other. Alas, Magic Mary and her magic coins are long gone—but many of the tablets remain, stored in various historical archives. And alas, no one can decipher the markings to tell which bias a given tablet corresponds to.

. . . or so we thought! But now bias-busting Bianca claims that she can decipher the markings and determine the coins' biases. How can we test her claim, given that *we* don't know how to decipher them?

Here's a good strategy. Go to an archive that contains a representative sample of tablets; draw a tablet at random; show her the markings-side, having her announce her guess as to whether it landed heads or tails along with her confidence in that guess; write down whether she got it right (but don't tell her); then draw a new tablet and repeat. Suppose we do this with many, many tablets, and then I tell you this: "Of the guesses she was 80% confident in, 79% were correct!" How confident are you now that Bianca can reliably recognize the 80%-biased coins—i.e. those that are 80% biased toward heads and those that are 80%-biased toward tails? Quite confident, I take it. For—in brief—it is rather surprising that so nearly 80% of those coins landed the way she guessed; and if she can reliably decipher them, that would explain why this is so. Conversely, if I instead told you that only *60%* of the judgments she was 80% confident in were correct, you should—for parallel reasons—suspect that she can*not* reliably decipher the markings of the 80%-biased coins, and instead that she is likely over-estimating the strength of these coins' biases.

Call this inference—from "Bianca was (mis)calibrated in her 80%-opinions" to "she probably can(not) reliably decipher the 80%-bias markings"—the **deciphered-to-right** inference, since it moves from her rates of being right to whether she has deciphered the markings. Clearly it is warranted in this simple scenario. And clearly there is an analogy between Bianca's test and Calvin's. If we can get clear on what exactly the analogy is and why the deciphered-to-right inference works for Bianca, it'll show us what needs to be the case for the rational-to-right inference to work for Calvin.

In fact, that's one of the main claims of this paper: if we want to know whether and to what extent we can expect the rational-to-right inference to work in a given scenario, imagine a parallel scenario for Bianca and her coins to see whether and to what extent the deciphered-to-right inference will work in that scenario.

So: Why does the deciphered-to-right inference work in this scenario? I said that it is because the hypothesis that she can(not) decipher the coins would help explain her calibration if she is (mis)calibrated. But what does that mean more precisely, and why is it true?

What it means more precisely is this. Before I tell you about Bianca's calibration, you should think to yourself:

> "If she can reliably recognize the 80%-biased coins, then the coins she says '80%' on will be (on average) around 80%-biased in the way she predicts—and conditional on *that*, I'm confident that roughly 80% of those tosses will land the way she predicts. Meanwhile, if she *can't* reliably recognize whether a coin is 80% biased, it's much more likely that a different proportion will land the way she predicts—for example, if she's over-estimating the bias, probably only 70% or 60% of the coins she says '80%' on will land the way she predicts."

Thus the evidence you received—that 79% of her 80%-opinions were correct—is much more likely given that she can decipher the 80%-biased coins than it is given that she cannot; so it provides reason to think she can do so. Conversely, if you learn that only 60% of her 80%-opinions were correct, this is much more likely given that she's over-estimating the bias of the coins, so it provides reason to think that she is over-estimating.

The driving force of the deciphered-to-right inference, then, is that hypotheses about whether she is deciphering the coins' biases, over-estimating them, or under-estimating them, each have direct and strong implications for how many of the coins *you* should expect to land the way she guesses.

Crucial question: why is this so? Answer: because hypotheses about the (average) biases of groups of coins have two very specific effects on how confident you should be in the outcomes of their tosses. First, you should **defer** to the average biases of the coins in setting your opinion for how a given coin will land: conditional on the coins corresponding to Bianca's 80%-opinions having an average bias of $x\%$ toward her predictions, you should be $x\%$-confident that each of those predictions will be true. Second, this deference is **independent**: regardless of how her other predictions turn out, it is still the case that conditional on the coins having an average bias of $x\%$ toward her prediction, you should be $x\%$-confident that her next prediction will be true.[16] Combined, these principles drive the deciphered-to-right inference by making it so that conditional on the coins having an average of $x\%$ bias toward Bianca's predictions, you're confident that roughly $x\%$ are true.

Upshot: for the rational-to-right inference to work in Calvin's case, analogous deference and independence principles must hold. What does the analogy amount to?

Bianca takes a bias-deciphering test in which she announces her best guesses about how coins with various markings landed, along with her confidence in those guesses. We want to use her resulting calibration score to draw conclusions about whether she is reliably deciphering the coins' biases, or over-estimating them, or under-estimating them. Meanwhile, Calvin takes a calibration test on which he announces his best guesses about the true answers to binary questions of various kinds, along with his confidence in those guesses. We want to use his resulting calibration score to draw conclusions about whether he is rational, overconfident, or underconfident.

For each tablet Bianca is shown, there is a fact about what the corresponding coin's bias was. Likewise, for each question Calvin assesses, there is a fact about the rational degree of confidence he should have in the possible answers.

We wanted to know whether Bianca can tell what the markings mean for the biases of the various coins. Likewise, we want to know whether Calvin can tell what his evidence means for the rational degree of confidence he should have in the various answers.

In Bianca's case, the deciphered-to-right inference went though because we should defer to the *biases* of the coins, and do so independently of how her other predictions turn out. Likewise, then, in Calvin's case: the rational-to-right inference will go through when and because we should defer to the *rational* degrees of confidence for Calvin to have in his answers, and do so independently of whether his other answers turn out to be true or false.

What does this mean more precisely? Consider all of the guesses Calvin assigns 80% confidence to—his 80%-opinions. Label them $q_1, ..., q_n$, so q_i is the claim that *the ith claim that Calvin was* 80% *confident in on this test (whatever it is) is true.*[17] We can entertain different hypotheses for what the average *rational* confidence is for Calvin to have in these claims. Let $\overline{\boldsymbol{R}}$ be this quantity, whatever it is.[18] Perhaps Calvin's

[16] In standard setups of our case, these two principles follow from the well-known Principal Principle and its refinements (Lewis 1980, 1994; Hall 1994; Briggs 2009b). See below for formal statements of their analogues in Calvin's case.

[17] For simplicity I assume you know that there are n such opinions. To generalize to the case where you don't know how many there are, we need to assume that learning how many there are would not affect our Deference and Independence principles below, and would not affect your confidence in what the (average) rational opinion for Calvin to have is. The inference will then go through by performing the reasoning described below, averaging over the various values n might take.

[18] Formally, let $R(q_i)$ be the *R*ational confidence for Calvin to have in any given claim q_i. Then $\overline{R} := \frac{1}{n}\sum_{i=1}^{n} R(q_i)$. I'll assume that the opinions that are rational for any given person can be modeled with a precise probability function. The same sort of reasoning may go through if the rational degrees of confidence were not unique (Schoenfield 2014) or not precise (Schoenfield 2012); for discussion of the

80%-opinions are on average rational, in which case this quantity will be 80%: $\overline{R} = 0.8$. Or perhaps they are on average overconfident (or underconfident), in which case it will be lower (or higher) than 80%: $\overline{R} < 0.8$ (or $\overline{R} > 0.8$).

Let q_i be any of Calvin's 80%-opinions. If you learn what the average rational opinion for Calvin to have in those opinions is, how does that affect your opinion in q_i? For the case to be analogous to Bianca's, you must defer. Let $\boldsymbol{P}$ be a probability function representing *your* rational degrees of confidence. Then what we need is:

> **Deference:** Upon learning that the average rational confidence for Calvin to have in his 80%-opinions is x%, you should be x% confident in each of them.
> For all q_i: $P(q_i|\overline{R} = x) = x$.

How plausible is Deference? As I'll come back to in the conclusion, that depends heavily on the epistemological theory we accept.[19] Importantly, no tenable epistemological theory will support a *stronger* deference principle than Deference—meaning that it's the tightest a connection between being rational and being right that we'll find. In particular, this means that all tenable epistemological theories will allow *at least* as much predictable rational deviations from calibration as those I illustrate in §§4–5.

Why think Deference, holds, even in this simple scenario? It deserves far more discussion, but let me say two things in its defense.

First, Deference tells you to defer to the opinions that are *rational* for Calvin to have, not the opinions he in fact has. Moreover, in our setup you don't know what claims are expressed by Calvin's 80%-opinions—q_i is simply the claim that *the ith claim on this test that Calvin was* 80%*-confident in (whatever that is) is true.* Thus you have virtually no evidence about the q_i. Meanwhile, Calvin has strictly more evidence than you about these claims—he knows all you do about the setup of the test, plus he knows *which* claims he was 80%-confident in, and therefore knows which facts bear on their truth. So conditional on Calvin's (more informed) evidence making it rational for him to be (on average) x% confident in these claims, it seems reasonable for you to be x% confident in it.

Second, there is a strong intuition that the rational-to-right inference is *sensible*: it in principle makes sense to run calibration studies to test for overconfidence. As we'll see, whether this is so depends on whether a principle like Deference holds. Thus anyone

(de)merits of such models, see White (2005, 2009a); Schultheis (2018); Carr (2019).

[19] Deference is an interpersonal, rationalized, and "averaged" generalization of the well-known Reflection principle (van Fraassen 1984; Briggs 2009a; Christensen 2010b; Mahtani 2017). Appendix A.1 shows how this "averaged" version can be derived from a more familiar "point-wise" version. Whether *interpersonal* deference principles hold is highly dependent on the debate between uniqueness and permissivism (e.g. White 2005; Schoenfield 2014, 2019; Horowitz 2014b, 2019a; Greco and Hedden 2016; Schultheis 2018). Whether *rationalized* deference principles hold is highly dependent on debates around higher-order evidence (e.g. Williamson 2000, 2019; Christensen 2010b; Lasonen-Aarnio 2013, 2015, 2019; Elga 2013; Horowitz 2014a; Salow 2018; Dorst 2019a,b, 2020). Deference will be a theorem in our setup given uniqueness plus higher-order certainty; it'll be approximately true under a variety of weaker theories.

who thinks the rational-to-right inference makes sense in principle is under pressure to accept an epistemological theory that can support strong deference principles.

Deference explains why the rational-to-right inference fails in many of our initial cases (§2). You shouldn't defer to Rajat (Case 1) because you something he doesn't—namely, that he's a brain in a vat. Likewise for Georgie—you know she had a bad geography teacher (Case 2). Similarly, when I saw that my 60%-biased coin landed heads only 30 of 100 times, I had evidence that you didn't when you formed your (rational) opinions about how it would land, so I shouldn't defer to them. Similarly for our final case—I shouldn't defer to your opinion about q_i if I know that it's in the set $\mathcal{W}$ of guesses you were wrong about, since you (of course) didn't know you were wrong about them when you formed your guesses.

However, Deference doesn't explain why the rational-to-right inference fails in our case of the misprinted coins. In that case, I haven't yet drawn the coin from the urn, so I defer to your rational opinions, yet I know you'll be miscalibrated. What's missing?

This is where we need our second assumption to make Calvin's case analogous to Bianca's: independence. This says that once you learn the average rational confidence for Calvin to have in his 80%-opinions, learning about whether some of those opinions were true or false doesn't affect your confidence in the others. Precisely:

> **Independence:** Given that the average rational confidence for Calvin to have in his 80%-opinions is x%, further learning that certain of these opinions are true or false shouldn't affect your opinion in the others.
> For all $q_{i_0}, ..., q_{i_k}$: $P(q_{i_0}|\overline{R} = x, q_{i_1}, ..., q_{i_l}, \neg q_{i_{l+1}}, ..., \neg q_{i_k}) \;=\; P(q_{i_0}|\overline{R} = x)$

How plausible is Independence? Again, there is much more to be said, but it is well-motivated as a first approximation—after all, you know the test questions were selected randomly, so learning whether some are true or false shouldn't (significantly) affect your deference to information about Calvin's rational opinions on others.[20] Independence explains why the rational-to-right inference fails in the case of the misprinted coins—in that case, we know that if the first toss lands heads, then the rest of them will as well.

Deference and Independence imply that the rational-to-right inference is warranted in our simple scenario: learning that Calvin's 80%-opinions were (mis)calibrated provides strong evidence that they were (ir)rational. This is because the assumptions make the case analogous to Bianca's: "(average) rational confidence for Calvin" plays the same epistemic role for you as "(average) bias of Bianca's coins." Just as the deciphered-to-right inference goes through in Bianca's case because you should defer (independently) to the biases of the coins, likewise the rational-to-right inference will go through in Calvin's

[20] This is at best approximately true, as learning that *all* of Calvin's other 80%-opinions were false should make you suspect that the test is tricky. What's definitely true is that the q_i are *exchangeable* (order doesn't matter) given $\overline{R}$. Using this we could prove more general versions of the formula derived §A.2 by using beta-binomial distributions rather than binomial ones. The reasoning will be similar, and the closer the q_i come to being independent, the stronger the rational-to-right inference will be.

case when you should defer (independently) to the rational opinions for Calvin.

In particular, conditional on Calvin's 80%-opinions being on average rational, you should be quite confident that roughly 80% of them will be true; and conditional on his 80%-opinions being on average overconfident (say, the average rational confidence is 60%), you should be quite confident that less than 80% (roughy 60%) of them will be true. Therefore when you learn that a given proportion of these opinions are true, that provides you with strong evidence about what the (average) rational confidence for Calvin to have is—i.e. about whether his actual opinions are (on average) rational.

To give a simple example, suppose you are initially equally confident that the average rational confidence ($\overline{R}$) for him to have in his 80%-opinions is any of 60%, 61%,..., or 99%. Suppose there are 50 such opinions. Let's say he is *substantially overconfident* if the average rational confidence in his 80%-opinions is less than 75% ($\overline{R} < 0.75$). Then you are initially 37.5% ($\frac{15}{40}$) confident that he is substantially over-confident. But if you were to learn that 70% of those opinions were true, then the rational-to-right inference is warranted: your confidence that he's substantially overconfident should jump to 78%.[21]

Upshot: despite a variety of concerns, the rational-to-right inference can be put on a firm theoretical foundation: when Deference and Independence hold, it is warranted.

By the same token, however: when Deference *fails*, the exact same reasoning will show that the rational-to-right inference fails with it. For example, suppose that conditional on the average rational confidence being 80%, you should be *70%* confident in each of Calvin's 80%-opinions: $P(q_i|\overline{R}=0.8) = 0.7$. Then (if Independence holds) you should be confident that if Calvin's rational, 70% of his 80%-opinions will be true—and thus finding out that 70% of such opinions are true (he's slightly over-calibrated) will be evidence that he's *rational*, rather than overconfident!

Thus we arrive at the key result:

> **Deference is Key:** Given Independence, the tenability of the rational-to-right inference in a given scenario stands or falls with the tenability of Deference.

So the crucial question is: how robust is Deference to variations in our simple scenario? §4 argues that it is very fragile: there are common scenarios in which Deference systematically fails, and hence we should not expect rational people to be calibrated. However, §5 argues that these failures of Deference and the corresponding rational deviations from calibration are in principle predictable—meaning that a more nuanced type of calibration study is possible.

4 The Limits

The real world isn't like the simple scenario, for you know a whole lot more about the test: its content, how it was constructed, what the experimenters were trying to

[21] The general formula for this update is given in §A.2.

show, what sorts of subjects were involved, and so on. Each of these bits of information threatens to undermine Deference and Independence in certain situations—and exploring the contours of these threats is important for having a full theory of the rational-to-right inference. Here I'll focus on just one type of information that a calibration study inevitably provides: our subject's full calibration curve—and, therefore, their overall proportion of true answers. Call that proportion their **hit rate**. Does knowing the hit rate cause a problem for the rational-to-right inference?

Yes—the hit rate tells you which sorts of *rational deviations* from calibration to expect. To see why, start with a simple version of Bianca's case.[22] Suppose in our archive all the tablets come from one of two coins—one that is 60% biased towards heads, the other that's 90%. Suppose we know that Bianca *can* decipher the coins, and that we'll randomly choose a couple dozen tablets from the archive. Should you expect her to be calibrated? Yes—but you should also expect her hit rate to be around 75%, because (1) she'll always guess heads (every coin is biased in favor of heads over tails), (2) we expect roughly 90% of the 90%-biased coins to land heads and 60% of the 60%-biased coins to do so, and (3) we expect they'll be roughly a 50-50 split between these coins ($0.5 \cdot 0.6 + 0.5 \cdot 0.9 = 0.75$).

Now suppose it turns out that Bianca's hit rate is *below* 75%. Should you still expect her to be calibrated? Definitely not. This is easy to see in extreme cases: if the hit rate is *very* low (say, 50%), it's impossible for her to be calibrated—since the lowest credence she'll assign is 60%. Similarly if it's less extreme: whenever the hit-rate is below 75%, you should expect Bianca to be over-calibrated.

The reason is that the hit-rate information breaks your deference to the biases of the coins. The connection between the biases of the coins and the frequency with which they land heads is probabilistic and therefore *loose.* Thus learning that the coins landed heads less often than you'd expect provides evidence that this is one of the cases where the biases and the frequencies came apart. That means upon learning that the bias of a given toss was 60% (90%), you should temper your deference downwards and be *less* than 60% (90%) confident that it landed heads. (Similar lessons apply if it turns out Bianca has a *high* hit rate: you should expect her to be *under*-calibrated.)

The same lesson applies to Calvin: when you learn that his hit rate on some set of questions was low (or high), this provides evidence that it was one of the scenarios in which there is a gap between being rational and being right—that fewer (or more) of Calvin's guesses were correct than he'd be rational to expect. Thus you should temper your deference downwards (or upwards): conditional on the average rational confidence in his answers being x%, you should be less than (more than) x% confident in a given answer.

Now let's state this line of reasoning more carefully. Let's assume that Calvin's hit

[22] Thanks to Daniel Rothschild for putting his finger on this way of explaining the problem.

rate, H, will be (approximately) equal to the *rational* hit rate, H_r—i.e. the hit rate he'd have if his degrees of confidence were rational. Since when faced with the question "A or B?" Calvin will guess the option he's more confident in, and he *should* guess the option he *should* be more confident in, this amounts to the assumption that in such binary-choice questions, Calvin will (usually) be more confident of A than B iff he should be. Grant this assumption for now—§4.1 explains why it's a reasonable one.

Granting the hypothesis that $H = H_r$, we can see that the rational-to-right inference will break when we learn Calvin's hit rate because this will break Deference. Consider again whether Calvin's 80%-opinions are rational. Learning Calvin's hit rate does not itself significantly affect your opinion this question—after all, learning (merely) the *rational* hit rate shouldn't affect your opinion whether his 80% opinions are rational, and we're granting that his hit rate equals his rational hit rate.[23]

So learning his hit rate doesn't shift your opinions in his rationality. But it *does* shift your opinions in the truth-values of his answers—for example, if his hit rate is low, you know many of his answer are wrong. This gives you information that he couldn't have had when he formed his opinions (he can't know how many of his opinions are right when he's in the process of forming them). Therefore this shift in your opinions about the truth-values should temper your deference to his rational credences.

For instance, suppose you learn that Calvin's hit rate is abnormally low—say, 50%. (75% is normal, since it's the average of 50–100%.) Now suppose you learn that Calvin's 80%-opinions were on average rational—should you be 80% confident in each of them? No! You should be *less* confident than that, since you know that more of them are false than he (rationally) expected. Thus although absent any information about the test you defer to his rational opinions, given hit-rate information you don't:

$$P(q_i|\overline{R} = 0.8) = 0.8, \quad \text{but}$$
$$P(q_i|\overline{R} = 0.8, H = 0.5) < 0.8$$

Thus conditional on his 80% opinions being rational, you should only be (say) 70% confident in each one being true. And conditional on his 80% opinions being *over*confident, you should be even less confident—say, 60%—in each one being true. If so, then—by exactly parallel reasoning to that at the end of §3—learning that only 70% of his 80%-opinions are true (he's slightly over-calibrated) is evidence that he's *rational.* For it's evidence that his confidence matched his accuracy as as could be expected, given the difficulty of the questions. The rational-to-right inference is inverted. (Likewise, if you learn that Calvin's hit rate is abnormally *high*, the inference will be inverted in the other way—learning that he's slightly *under*-calibrated will be evidence that he's rational.)

Here's a simple example. Again suppose you are initially equally confident that

[23] Precisely: for any t, s, $P(\overline{R} = s|H = t) \approx P(\overline{R} = s|H_r = t) = P(\overline{R} = s)$.

the average rational confidence for him to have in his 80%-opinions ($\overline{R}$) is any of $60\%, 61\%, ..., 99\%$, and there are 50 of them. Suppose that learning that his hit rate was 50% does not affect your confidence in any of these hypotheses, but it has the effect of tempering your deference in each downward by 10%: $P(q_i|\overline{R} = x, H = 0.5) = x - 0.1$ (so, for example, if his 80%-opinions are rational you should be 70% confident in each of them). Say that Calvin is *approximately rational* if $0.75 \leq \overline{R} \leq 0.85$. Then you are initially 27.5% ($\frac{11}{40}$) confident that he's approximately rational, but upon learning that 70% of his 80%-opinions are true (he's slightly over-calibrated), you should *increase* this confidence to 61%. Meanwhile, you should *decrease* your confidence that Calvin is substantially overconfident ($\overline{R} < 0.75$) from 37.5% to 22%, inverting the effect from the end of §3.

In summary, we've arrived at the following result:

> **Hit Rates are Key:** The rational-to-right inference works only when (rational) hit rates are moderate—on any set of questions on which (rational) hit rates are high (or low), rational deviations from calibration should be expected.

This qualitative claim raises a quantitative question: *how much* deviation from calibration should we expect, as hit rates vary? In §5 I'll show how we can answer this question under the assumption that people's actual hit rates match the rational ones; so first, we need to clarify why this is often a reasonable assumption.

4.1 (Rational) Hit Rates

To do so, we need to get a bit clearer on what the overconfidence hypothesis and its alternatives might be (§2). We've been simplifying by focusing on the "overconfidence effect"—in fact, many studies find wildly different calibration curves for different types of questions. Sometimes people are over-calibrated at all levels of confidence; other times they are over-calibrated at high levels of confidence and under-calibrated at low levels of confidence; other times they are under-calibrated at all levels of confidence, and so on (more on this in §5; see Koehler et al. 2002; Brenner et al. 2005). Translating these calibration curves to corresponding (ir)rationality hypotheses, the varying types of possibilities are shown in Figure **??**. In this figure, interpret the lines as averages: for example, the "over-extreme" hypothesis says that when a person's actual confidence is 80%, the confidence it is on average rational for them to have is merely 60% (as indicated by the red dot).

With these options on the table, the live (ir)rationality hypotheses are claims of the form, "For questions of type X, people's confidence obeys (ir)rationality hypothesis Y", where X is some specification of question-type, and Y is a curve having a shape like those in Figure **??** (Brenner et al. 2005). For example, *ecological models* have proposed that if X is "questions sampled randomly from a natural domain," then Y is the rational

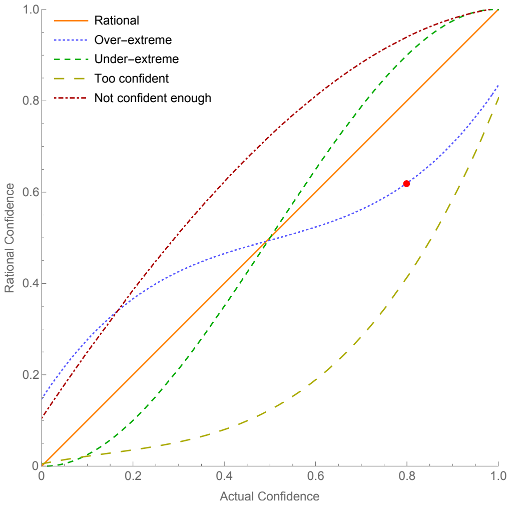

Figure 3: The Various (Ir)rationality Hypotheses

curve (Gigerenzer et al. 1991; Juslin 1994); meanwhile, *case-based judgment models* have proposed (among other things) that if X is "questions on which case-specific evidence is statistically weak and the base rate of truths is moderate," then Y is the over-extreme curve (Griffin and Tversky 1992; Koehler et al. 2002; Brenner et al. 2005).

Here is an important prediction of any such (ir)rationality hypothesis: if the alternative claims someone is guessing between are from the same domain, then *people's guesses will tend to be rational.*

Why? Note that all proposed (ir)rationality hypotheses have positive slopes, meaning that (on average) higher rational degrees of confidence correspond to higher actual degrees of confidence. Take any such (ir)rationality hypothesis, and consider a guess between a set of claims that it treats as in the same domain—say, "Which is bigger: Rome or Madrid?" The rational guess for Calvin is the one that he should assign higher confidence to: if $R(Rome) > R(Madrid)$, it's rational for Calvin to guess *Rome*; and if $R(Rome) < R(Madrid)$, it's rational for Calvin to guess *Madrid*. But since higher rational degrees of confidence correspond to higher actual degrees of confidence, the (ir)rationality hypothesis predicts that if the former, then Calvin's *actual* confidence will be higher in *Rome* than in *Madrid*—i.e. he'll guess *Rome*; and if the latter, Calvin's actual confidence will be higher in *Madrid*—i.e. he'll guess *Madrid*. Either way, the (ir)rationality hypothesis predicts that Calvin will guess rationally.[24]

[24] Formally, let $C(q)$ be Calvin's actual confidence in q, and let an (ir)rationality hypothesis be a function $f : [0,1] \to [0,1]$ mapping actual degrees of confidence to (average) rational degrees of confidence: $R(q) = f(C(q))$. Any such function that is monotonically increasing ($f(x) > f(y)$ iff $x > y$) will be such that if $R(q) > R(p)$, then $f(C(q)) > f(C(p))$, hence $C(q) > C(p)$. Notably, since f is most plausibly interpreted as an average, there will be exceptions to this connection between rational and actual guesses. How common such exceptions will be depends on (1) how steep the slope of the (ir)rationality hypothesis is, and (2) how widely the deviations from f are distributed. Notably, if we

Upshot: at least when people are guessing amongst claims that come from the same domain, all (ir)rationality hypotheses will agree that people's guesses will tend to be rational. People's hit rate is fully determined by their guesses—so if their hit rate is low, then (since their guesses will tend to be rational) this means that the *rational* hit rate is low as well. In other words, in many studies it is common ground amongst all (ir)rationality hypotheses that we are in a situation in which we know that Calvin's hit rate will be (close to) the rational hit rate. Since Hit Rates are Key, this means that when we learn the hit rate on our study, the rational-to-right inference will fail in predictable ways.

5 The Implications

At this stage we've established that we should not expect rational opinions to be calibrated on sets of questions for which the hit rates turn out to be low (or high)—even if the questions were selected randomly from a domain that is representative of people's knowledge. This means that no matter how carefully we construct our test, we cannot evaluate the overconfidence hypothesis by simply checking whether people's opinions are calibrated—for we should often not expect *rational* opinions to be calibrated.

What should we do, instead? My proposal is that we use the Bianca analogy to predict the rational *deviations* from calibration given our test setup, and then compare observed calibration curves to those predictions. We can do this in three steps:

1) Choose a test-construction procedure, along with a hypothesis about how this procedure will sample from rational opinions and right opinions.
2) Translate that hypothesis into the Bianca analogy and use it to build a simulation of the rational opinions.
3) Compare the predicted (mis)calibration of the rational opinions from this simulation to the actual calibration curves we observe.

I'll spend the rest of this paper illustrating how this methodology can work and arguing that it calls into question the standard interpretation of certain empirical effects.[25]

5.1 The Hard-Easy Effect

It turns out that the "overconfidence effect" is an overgeneralization: it is not the case that people are in general over-calibrated on binary-question tests. Rather, we can distinguish the tests that are *hard* from those that are *easy* based on the hit rate: an

use the *average* hit rate (across subjects) on a test, and assume that subjects share similar evidence, such deviations from rationality should cancel out, and the average actual hit rate should be quite close to the average rational one.

[25] This methodology is a generalization of the simulation-based approaches found in, for example, Juslin et al. (1997, 1999); see §2 for more on the relation between the two.

easy test is one with a hit rate of at least 75%; a hard test is one with a hit rate of less than 75%. The empirical generalization that subsumes the "overconfidence effect" is called the **hard-easy effect:** people tend to be over-calibrated on hard tests and *under*-calibrated on easy tests—see Figure 3. (The reason we see the "overconfidence effect" on general-knowledge trivia tests is simply that most such tests turn out to be hard.) The hard-easy effect has been called "fundamental bias in general-knowledge calibration" (Koehler et al. 2002, 687), and is widely cited as one of the core pieces of evidence in favor of the overconfidence hypothesis (e.g. Lichtenstein et al. 1982; Keren 1987; Gigerenzer et al. 1991; Griffin and Tversky 1992; Juslin 1994; Juslin et al. 2000; Koehler et al. 2002; Brenner et al. 2005; Hoffrage 2004; Moore and Healy 2008; Glaser and Weber 2010). The standard interpretation is that people do not make sufficient adjustments for task difficulty, leading them to be overconfident on hard tests and underconfident on easy ones.[26]

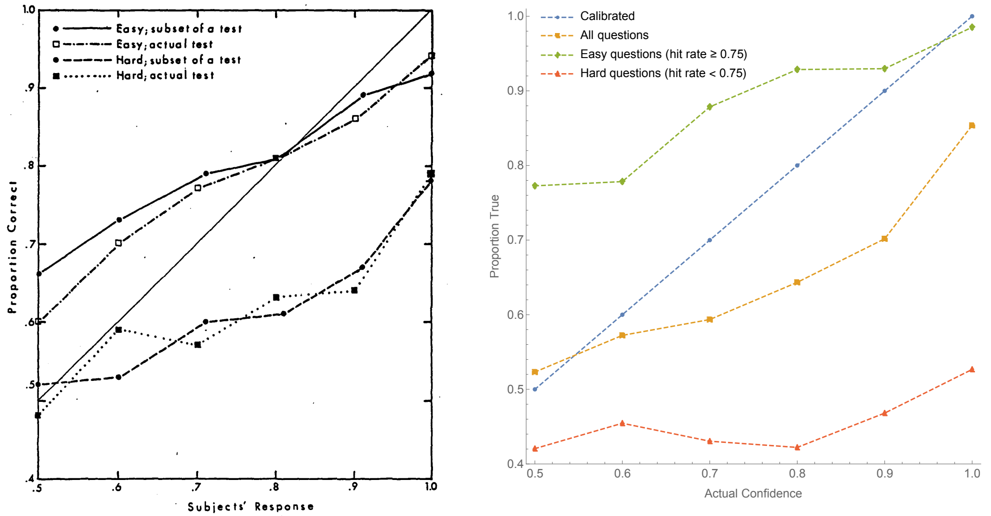

Figure 4: The hard-easy effect. In both graphs, top curves are easy sets of questions; bottom curves are hard ones. **Left:** Lichtenstein et al. (1982). **Right:** My study.

The hard-easy effect is one of the core pieces of evidence offered in favor of various versions of the irrationality hypothesis. However, we now know that systematic patterns of miscalibration should sometimes be expected of *rational* people when the hit rate varies. The question, then, is whether these empirical effects should be surprising given the null hypothesis that people are (approximately) rational.

[26] A closely related effect is the **base rate effect**: on tests in which subjects are simply presented with a series of claims and then rate their confidence from $0-100\%$, the overall proportion of truths (the base rate) has a dramatic effect on people's calibration curves (Lichtenstein et al. 1982; Koehler et al. 2002): low base-rates tend to lead to over-calibration and high base-rates tend to lead to under-calibration. Though for brevity I will omit simulations for this effect, the methodology and predictions are exactly the same, since learning the base rate on a set of questions breaks our Deference condition in the exact same way that learning the hit rate does (§4).

5.2 Testing for Rationality

How can we know what to expect rational calibration curves to look like on tests of various types? The way we're going to answer this question is by returning to our coin analogy with Bianca. We are now going to assume that she *can* (at least usually) decipher the tablet markings—and thus set her confidence (at least approximately) equal to the biases of the coins—and go on to simulate what calibration curves we should expect from her as we vary the method of constructing the test and the difficulty of various sets of questions from it.

Step 1 is to choose our test-construction procedure, and form a hypothesis about how this procedure will sample from the rational opinions and the right opinions. In particular: (i) how likely are we to include a question on which the rational credence in the answer is 50%? 60%? Etc. (ii) And on any given test we give, do we defer to the rational opinions? If so, how robust is that deference—does Independence hold, or would learning of false (true) answers temper our deference downwards (upwards), away from the rational credence? These questions matter because they affect (i) how often our simulations present Bianca with coins of various biases, and (ii) how robustly the bias of the coins lines up with our expectations about how many of them land heads.

First focus on the simplest case: a test on which we can reasonably suppose that (i) the questions we pull are *equally* likely to have any level of rational confidence in their guess, between $50-100\%$; and on which (ii) our deference is quite robust.

One way to try to form such a test is to make one on which we pull questions randomly from a well-defined, representative domain on which we can expect that the accuracy of people's evidence will not be systematically correlated across questions.

This turns out to be a difficult criterion to meet, but I'll take a standard paradigm from the literature (Gigerenzer et al. 1991), and pull pairs of American cities randomly from the top-20 most populous cities, and ask people to guess which they has a bigger population and to rate their confidence in that guess.

On a representative-question test like this, it's reasonable to posit that the rational credences in answers will be fairly uniformly distributed between $50-100\%$. How robust your deference should be is a more vexed question—if we discover Calvin is wrong about whether San Francisco is larger than Phoenix, should that temper our deference to his rational opinion about whether San Jose is bigger than Austin? Perhaps—but let's ignore that for now (and come back to it in a moment).

Given this, we can model perform Step 2: model (and then simulate) our test using the Bianca analogy. We toss a number of coins (equal to the number of questions on our test), selecting them uniformly at random from coins of varying biases between $50-100\%$[27], have her guess how they'll land and rate her confidence in that guess, and

[27] I simplify by tossing coins of biases $50-100\%$ and having her always announce heads, rather than tossing coins of biases between $0-100\%$ and having her first guess whether the coin lands heads or

record her calibration curve. This is a single trial. Repeat this procedure thousands of times, and now look at the average results on trials (sets of questions) that have various hit rates. What do we expect to see for question-sets of various hit-rates?

For all simulations, I'll display two versions. The **perfection model** assumes Bianca always gets the biases of the coins exactly right (analogy: Calvin is always perfectly rational). The **noise model** assumes that Bianca's announced confidence is a random perturbation of the bias of the coin—capturing the idea that she may be a reliable but imperfect at deciphering the coins' biases (analogy: Calvin's confidence may be a reliable but imperfect tracker of the degree of confidence his evidence warrants).[28] The most plausible versions of the rationality hypotheses are ones in which there is some such error—though of course it's worth emphasizing that whenever their *is* such error, the person by hypothesis is not fully rational (cf. Brenner 2000). Nevertheless, since such deviations from rationality will be randomly distributed, there is still a good sense in which people who conform to such models are approximately rational.

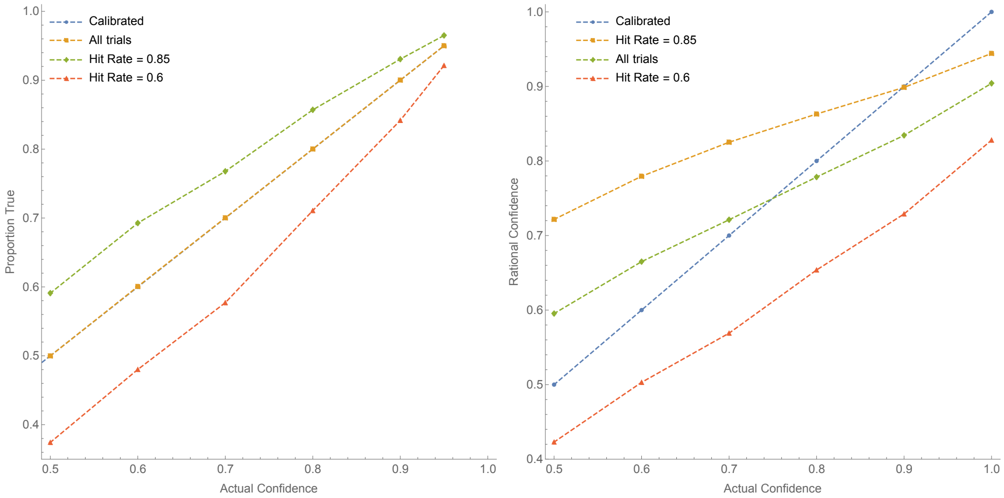

Figure 5: Random tests, restricted to various hit rates. **Left:** Perfection model. **Right:** Noise model (100,000 trials each).

For illustration, the expected calibration curves for Bianca at various hit rates are displayed in Figure 4. When we consider all trials together, Bianca is calibrated—

tails. The underlying statistics are the same.

[28] I assume the errors are normally distributed with mean 0; Figure 4 uses standard deviation 0.2. This model takes inspiration from "error models" (Erev et al. 1994; Pfeifer 1994; Juslin et al. 1997, 1999), but the interpretation is importantly different. Their models treats people's reported opinions as imperfect indicators of their *true* opinions (or, in some variations, the objective frequencies of truths), whereas mine treats people's reported opinions as imperfect indicators of the *rational* opinions. It is plausible that the latter errors will be larger than the former (it's harder to know what you *should* think than to know what you *do* think!). Moreover, while tests of the variance of people's reports have suggested that error in reporting their true confidence cannot account for the observed miscalibration (Budescu et al. 1997), these tests cannot test for error in matching their reported confidence to the rational confidence.

perfectly so in the perfection model; slightly less so in the noise model due to "scale-end effects" (Juslin et al. 2000)—at the end-points of the confidence scale, errors can only go in one direction, resulting in the tilting of the curve. But amongst tests where the hit rate is low (high), Bianca tends to be over-(under-)calibrated—just as observed empirically with the hard-easy effect.

Why? Consider a given trial on which the proportion of heads was lower than usual. Why was it lower? One explanation is that this trial had an abnormally large proportion of coins that were biased against landing heads. A different explanation is that this test was unusual in the sense that more of the coins landed tails than you'd usually expect, given their biases. Both are likely to play a role in any given trial with a low hit rate. Bianca will account for the first factor in setting her degrees of confidence, since she can recognize the coins and see that more of them than usual have a low bias—but it is impossible for her to account for the second factor. The result? As we consider cases with more extreme hit-rates, Bianca will become increasingly miscalibrated. For example, take the perfection model—where Bianca is as sensitive to the biases of the coins as she could possibly be. On the binary-question test, on trials with a hit-rate of 75%, Bianca's average confidence was 75%; on trials with a hit rate of 90%, her average confidence is 77% (becoming under-calibrated); and on trials with a hit rate of 60%, her average confidence is 72.7% (becoming over-calibrated).

Upshot: even if the calibration tests contain questions that are random samples of the overall distribution of rational opinions (the best-case-scenario for the rational-to-right inference, as seen in §3), we would still expect some form of the hard-easy to emerge for rational subjects. Moreover, if they are merely approximately rational (the noise model), we should expect rational calibration curves that are qualitatively similar to the curves we observe empirically (compare the right side of Figure 4 to Figure 3).

Let's now perform Step 3 and apply this model to *my* study (pre-registration available here). I generated all pairs from the 20 most-populous U.S. cities, and recruited 200 U.S. residents through Prolific (90 F, 107 M, 3 Other; mean age 34.7). After giving them standard instructions about how to use the 50–100% confidence scale ("Ideally, 8 out of 10 of the things you're '80%' confident in should be true", etc.), I presented each with 21 pairs—20 randomly selected from the 190 pairs of the top-20 U.S. cities, and 1 attention check. (Data from those who incorrectly answered this check were excluded; only 1 participant failed.)

I pooled subjects' answers, and divided the questions into those that were *easy* (more than 75% of answers correct) and those that were *hard* (less than 75% correct). Figure 3 (page 20) above shows the calibration curves from my study—overall, amongst the hard questions, and amongst easy ones. The hard-easy effect was observed as expected—though it was especially stark. Amongst hard questions the average confidence was

75.1%, while the proportion true was only 45.2%.[29] Meanwhile amongst easy questions, the average confidence was only 84.7%, while the proportion true was 92.1%.[30] Somewhat unexpectedly, the test overall was slightly hard, with an average confidence of 79.8% and a proportion true of only 68.0%—accounting for the over-calibration observed across all questions in Figure 3.[31]

We can compare these results to both the perfection-model and noise-model predictions. As pre-registered, I generated these simulations by setting the number of questions Bianca faces to the size of the easy/hard/all-questions set, simulating millions of trials, and then removing trials with high/low hit rates until the mean hit rate matched the actual hit rate in the easy/hard/all-questions sets.[32] The perfection model has no free parameters; its comparisons are displayed on the left of Figure 5. As can be seen, each predicted curve crosses the actual curve but the overall- and hard-curves have significantly steeper slopes.

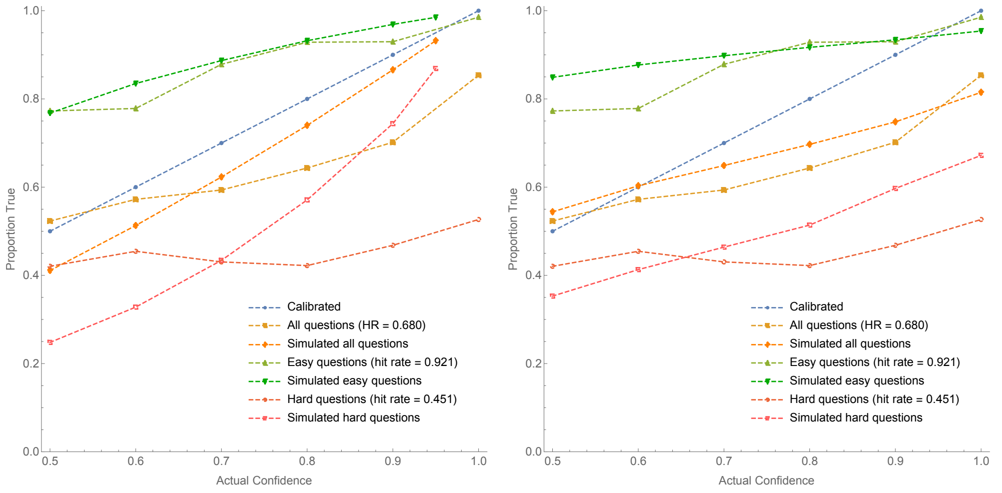

Figure 6: Random tests run with the observed hit-rates in my study. **Left:** Perfection model (5 million trials). **Right:** Noise model (8 million trials, noise parameter = 0.3).

The noise model has a free parameter for the standard deviation in the noise that leads people's credences to diverge from the rational credences. As pre-registered, to set

[29] The difference is significant: average confidence in hard questions ($M = 0.751$, $SD = 0.165$) was above proportion true of hard questions ($M = 0.452$, $SD = 0.498$), with a one-sided independent samples t-test of $t(2487) = 25.82, p < 0.001$, and $d = 0.808$.

[30] The difference is significant: average confidence in easy questions ($M = 0.847$, $SD = 0.162$) was below proportion true of hard questions ($M = 0.921$, $SD = 0.270$), with a one-sided independent samples t-test of $t(3156) = 10.37, p < 0.001$, and $d = 0.333$.

[31] Since unexpected, this test was not pre-registered; but the difference was significant: mean of confidence across all questions ($M = 0.798$, $SD = 0.170$) differed from mean truth-value ($M = 0.680$, $SD = 0.467$), with a two-sided independent samples t-test of $t(5013) = 14.97$, $p < 0.001$, $d = 0.336$.

[32] The main limitation was that the observed hit rates were too extreme for me to obtain enough trials with the observed hit rate using the actual number of test questions in each set, so I had to use a lower number of 60 questions in each easy/hard/all category. The shape of the generated curves is quite robust to this parameter.

this parameter I ran versions of the simulations with the parameter varying from $0 - 0.3$, and chose the one with the resulting predicted calibration curves that minimized mean squared divergence between the model predictions (amongst hard and easy subsets) and the actual curves. This set the noise parameter to 0.3, and the resulting comparison of curves is displayed on the right of Figure 5.[33] As can be seen, the predictions generated from the rational-credence-plus-noise model, though not a perfect fit, are generally close to the observed mis-calibration.

But this isn't the end of the story. One puzzling thing about the above simulations is why it was so difficult to find instances with hit rates as extreme as we observed in the real study (of 8 million trials, only 183 had hit rates at or below 0.515, while my study's hard questions had a hit rate of 0.452). A natural answer is the following: in tests that share a *common subject-matter*—such as my city-comparison tests, and many others[34]—we need to revise our assumption of Independence, since the subject's evidence will be highly correlated across questions. In particular, though we should expect that the opinions warranted by their evidence will *on the whole* be calibrated, we should also expect that a there will be random fluctuations in how calibrated they are across subject-matters. For instance, in my city-comparison test, some subjects will have evidence that warrants misleadingly strong opinions (only 70% of the opinions they should be 80% confident in are true), while others will have evidence that warrants misleadingly weak opinions (90% of the opinions they should be 80% confident in are true). Moreover, we expect these fluctuations in evidence to be correlated for a given person on a given subject-matter—if only 50% of the opinions Calvin ought to be 60% confident in on my test are true, we should expect that (say) only 60% of the ones he ought to be 70% confident in are true.

Here's a natural way to model this. Again there is a random number of coins of varying biases that Bianca can recognize, but this time there is random variation across tablet archives in how representative they are of the broader distribution of tablets—some archives have higher proportions of heads from a coin of a given bias than would be expected; other have lower proportions. Thus for each trial (visit to an archive), we generate a random misleadingness parameter and add it to the coin biases to determine how far the proportions of heads in this archive deviates from the biases of the coins.[35]

Although I had constructed these models before running my city-calibration test, it only occurred to me that they were an apt model of it after running the test and seeing how extreme the variation in hit-rates were. As a result, these comparisons were not

[33] Obviously these are not the most rigorous statistical methods, but they suffice to illustrate the conceptual points of this paper. It should also be noted that this is a rather high noise parameter, corresponding to a fair amount of random deviation of actual confidence from rational confidence.

[34] (Dunning et al. 1990; Vallone et al. 1990; Brenner et al. 1996; Koehler et al. 2002; Brenner et al. 2005; Hoffrage 2004; Glaser and Weber 2007; Merkle and Weber 2011; Brenner et al. 2012).

[35] In the displayed simulations this parameter is normally distributed with mean 0 and (for illustration) standard deviation 0.2. In these simulations I assume that the variation in misleadingness is only in the magnitude—not the direction—of the evidence, so it never pushes below 50%.

pre-registered, and therefore should be taken with a grain of salt. But it turns out to be *much* easier to find hit-rates as extreme as the ones we observed using this model, lending it some support. Running the same analysis as above yields find the optimal noise parameter at 0.15, and yields the comparisons in Figure 6.

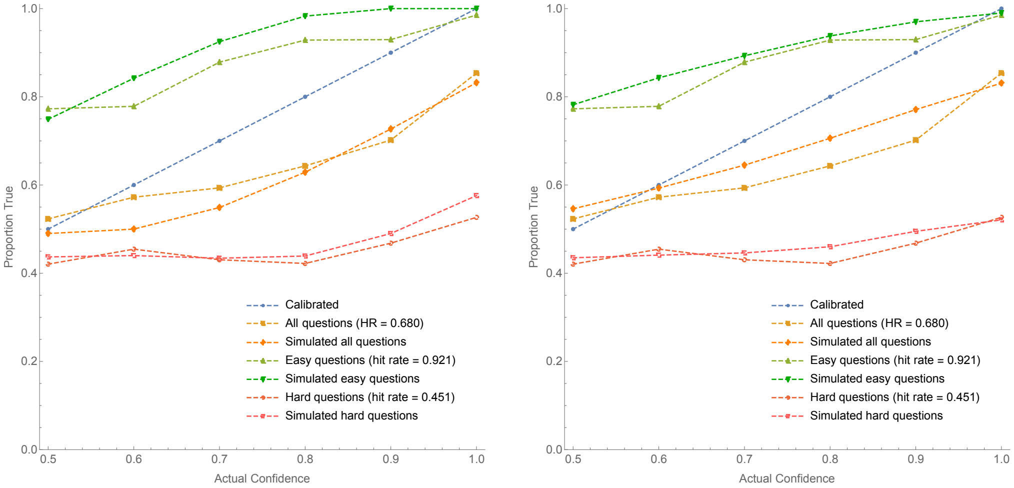

Figure 7: Tests with random misleadingness (parameter = 0.2), run using the observed hit rates in my study (20,000 trials). **Left:** Perfection model. **Right:** Noise model, parameter = 0.15.

The point? The rational models are by no means a perfect fit. However, the large-scale qualitative effects on this sort of study—such as the hard-easy effect, and the prediction that the subjects will be overall over-calibrated if the hit rate is below 75%—*are* predicted. Moreover, when we incorporate the possibility of random noise in subject's judgments *or* random misleadingness in subject's evidence, the observed calibration curves are close to what we should expect from rational people.

The important takeaway is that it is the (much smaller) deviations from *these* predicted curves that we must study systematically, not the deviations between people's actual confidence and the perfectly-calibrated line.

6 The Open Questions

I'll close by briefly considering a few open questions for the theory of rational (mis)calibration developed here: the philosophical tenability of Deference (§6.1), and the bearing of these results on the Dunning-Kruger (§6.2) and "over-precision" effects (§6.3). Readers uninterested in these issues may skip to §7.

6.1 The Tenability of Deference

I have built a theory of rational (mis)calibration—of the connection between being rational and being right—based on Deference. But, as mentioned in §3, Deference is the strongest tenable interpersonal deference principle—and there are many philosophical reasons to be worried about it. First, if epistemic rationality is *permissive*, then you may have different epistemic standards than Calvin (White 2005; Schoenfield 2014, 2019). If so, the fact that *his* standards rationalize being 80%-confident in q doesn't imply that *your* standards do—perhaps your standards warrant being systematically more cautious (less extreme) in your opinions than Calvin's do. If so, Deference will fail—for instance, you may temper your deference downward from Calvin even without hit-rate information: $P(q_i|\overline{R} = 0.8) < 0.8$. This sheds doubt on whether permissive views of epistemic rationality can justify the rational-to-right inference even in our best-case scenario (§3).

Similarly, if epistemic rationality is *modest*—meaning it can be rational to be unsure of what credences are rational—then it Deference must sometimes fail (Christensen 2010b; Elga 2013; Dorst 2019a). The only deference principle I know of that is both tenable in general in the case of modesty, and would warrant a variant of the reasoning from §3 is (a generalization of) the "Trust" principle in Dorst (2019a).[36] So far as I know, every other proposed principle (Elga 2013; Pettigrew and Titelbaum 2014; Gallow 2019)—or argument that there can be no such general deference principle (Lasonen-Aarnio 2015; Williamson 2019)—would allow wild and systematic deviations between what you learn about Calvin's rational credence and the resulting credence you should adopt. As a result, they would not undergird the rational-to-right inference even in ideal cases.

These issues are pressing, since permissivism and modesty are both thought to be highly applicable to any notion of rationality that applies to human reasoners. As such, it's important to figure out whether such theories can explain the connection between being rational and being right in a way that could undergird the methodology of calibration studies—or whether such views are committed to the claim that none of these results provide us with evidence of overconfidence.

6.2 The Dunning-Kruger effect

The Dunning-Kruger effect (Kruger and Dunning 1999) is the finding that those who are comparatively unskilled in a given domain are also unable to accurately assess how comparatively unskilled they are. Precisely: the gap between a person's relative performance on a test (which percentage of test-takers did they outperform?) and their

[36] The "variant" reasoning requires looking not at whether the (average) rational credence is *exactly* t, but instead whether it is *at least* (or at most) t—which would require pooling people's judgments into categories "at least 50% confident", "at least 60% confident"; and so on.

estimate of this number grows as relative performance decreases. For example, those in the 50th percentile may estimate that they are in the 60th percentile, while those in the 20th percentile may estimate that they are in the 50th percentile. This finding is routinely chalked up to a cognitive bias—a failure of the metacognitive ability to assess how competent one is (Dunning 2012).

The theory developed here tells against this, and instead supports the theory from Moore and Healy (2008) and Jansen et al. (2018). As we've seen, for any set of rational opinions, there will be tests that are hard and easy for those opinions—in particular, that will lead to low or high hit rate (§2). We have also seen that in any such test, as the test gets harder for a rational person, they will become increasingly over-calibrated—meaning the gap between performance (actual hit rate) and estimated performance (average confidence, i.e. estimated hit rate) will grow (§5.2). Therefore, since even for rational people, the difficulty of a test will vary depending on their knowledge and skills, a straightforward consequence is that rational people who perform less well on a test will over-estimate their performance more than rational people who perform better. The Dunning-Krueger should be expected of them.[37]

6.3 Rational Over-precision?

A different type of calibration test asks people to state various confidence intervals for the true value of some unknown parameter, such as the length of the Amazon. Empirically, there is a sense in which people tend to be systematically more over-calibrated on tests like this (Juslin et al. 1999; Moore and Healy 2008; Glaser and Weber 2010; Ortoleva and Snowberg 2015; Moore et al. 2015)—what has come to be known as "over-precision" (Moore and Healy 2008). The theory developed here may help explain this.

First note that this test can be translated to our familiar format (Tversky and Kahneman 1974): your 90% confidence interval for the length of the Amazon is "1000 to 5000 miles" iff you are 95% confident in both "The Amazon is at Least 1000 miles long" (L), and "The Amazon is at Most 5000 miles long" (M); your confidence intervals tend to miss the true value too often iff you are over-calibrated on claims like L and M. Now, for someone to be calibrated in their interval estimates, items must fall outside the range of their 90%-confidence interval exactly 10% of the time. Yet studies regularly find "miss rates" as high as 50%, and almost never lower than 10% (Glaser and Weber 2010, 243). Is this evidence for a more robust form of overconfidence?

Not obviously. Note that for hard binary-question tests, it is standard to see less than 75% of people's 95%-opinions being true—in fact, our noise models from §5 predict at least that much over-calibration when hit rates are low (Figures 4, 5, and 6). Now, Calvin's 90%-confidence interval for the Amazon's length in miles is "1000 to 5000" iff

[37] Contra Merkle and Weber (2011)—whose response to Moore and Healy (2008) illicitly assumes that Bayesians will have priors that match the objective frequencies on the test.

he is 95% confident in both“The Amazon is at Least 1000 miles” (L) and “The Amazon is at Most 5000 miles” (M). Our credence that both L and M are true (his interval covers the true value) is our credence in the former, multiplied by our credence in the latter given the former: $P(L \wedge M) = P(L) \cdot P(M|L)$. As we've seen, given that Calvin's credence is 95% in each of these claims, if he's only approximately rational and the test his hard, we should be only 75% confident in each of them. That means that *if they were independent*, we'd expect his interval to cover the true value (L and M to both be true) about $0.75 \times 0.75 = 56.25\%$ of the time, leading to a miss rate of around 44%.

But note that they are (by definition) *not* independent: if L were false (the Amazon is less than 1000 miles), M would necessarily be true; hence learning that L is true necessarily lowers the probability of M: $P(M|L) < P(M) = 0.75$. Thus we should expect a hit rate for the conjunction $L \wedge M$ of *less* than 56.25%, and hence a miss rate of greater than 44%; hence 50% does not seem especially surprising for hard tests. Moreover, by parallel reasoning we should expect less-than-10% miss rates only if *more* than 95% of a person's 95%-opinions are true. Yet we've seen that (due to scale-end effects) this is virtually never the case (none of our binary-question noise models—even with easy tests—see such high rates).

Thus it seems an open question whether attention to rationality of over-calibration combined with the non-independence of the individual probability judgments that compose an interval estimate might shed new light on empirically observed over-precision.

7 The Conclusion

Many have taken the results of calibration studies to demonstrate that people tend to be systematically overconfident in a way that is both dire and preventable. I've argued that the theoretical foundations of this inference are shaky (§2), but that we can secure them by articulating a probabilistic connection between being rational and being right (§3). Yet though this supplies a foundation to such studies, it also reveals a flaw: no matter how well-designed the study, rational people should be expected to be miscalibrated in systematic ways (§4). Using these systematic deviations, I proposed a modification of the standard methodology: we must use hit rates and other information about our study to predict the rational deviations from calibration, and then compare people's performance to those predictions. I illustrated how this can be done, and argued that it may overturn the standard interpretation of robust empirical effects (§§5–6).

If even a portion of this discussion is correct, is suggests that certain debates in philosophy and psychology are much closer than has been realized. Psychologists have had a long, spirited debate about the bearing of empirical results (like those of calibration studies) on human (ir)rationality.[38] Yet most contemporary philosophical debates

[38] For classic statements of the “irrationalist” approach, see Tversky and Kahneman (1974, 1983);

about rationality have been relatively isolated from these issues.[39]

As we've seen, these debates needn't—and arguably *shouldn't*—be isolated. Whether and the extent to which we have empirical evidence for overconfidence depends on the connection between being rational and being right. The proper formulation of such a connection is directly dependent on philosophical debates about the proper formulation of deference principles—and, relatedly, about permissivism and epistemic modesty (§6.1). Thus these philosophical debates have a direct bearing on the proper interpretation of these empirical studies. Conversely, the methods and results from calibration studies are directly relevant to ongoing the philosophical debate about how to understand the connection between being rational and being right.[40] For instance, simulations like the ones I used in §5—based on the methods developed by psychologists (Erev et al. 1994; Pfeifer 1994; Juslin et al. 1997, 1999, 2000)—can be used to make precise predictions about the relation between rational confidence and accuracy.

In short: both psychologists and philosophers have been investigating rationality—but often from radically different directions, and without substantial discussions. We've seen that the questions, methods, and tools from these investigations can be tied together in surprising and fruitful ways. That raises an exciting question: If we bring these investigations closer together, what other ties might we find?[41]

Kahneman et al. (1982); Kahneman and Tversky (1996); Fine (2005); Ariely (2008); Hastie and Dawes (2009); Kahneman (2011b); Thaler (2015). For defenses of "rational" approaches see Anderson (1990); Gigerenzer (1991); Oaksford and Chater (1994, 2007); Tenenbaum and Griffiths (2006); Hahn and Oaksford (2007); Hahn and Harris (2014); Harris and Hahn (2011); Tenenbaum et al. (2011); Griffiths et al. (2012); Cushman (2018).

[39] Though in recent years there are an increasing number of exceptions, e.g. Cohen (1981); Stich (1985); Kelly (2004, 2008); Crupi et al. (2008); Fitelson and Hawthorne (2010); Koralus and Mascarenhas (2013); Nebel (2015); Icard (2017); Hedden (2018); Mandelbaum (2018); O'Connor and Weatherall (2018); Singer et al. (2019); Doody (2020); Quilty-Dunn (2020).

[40] Joyce (1998); Littlejohn (2012); Pettigrew (2016b); Schoenfield (2016b); Horowitz (2014b, 2019b); Comesaña (2020); Staffel (2020).

[41] Thanks to Lyle Brenner, Liam Kofi Bright, Thomas Byrne, Fiery Cushman, Chris Dorst, Dmitri Gallow, Cosmo Grant, Brian Hedden, Thomas Icard, Joshua Knobe, Harvey Lederman, Matt Mandelkern, Don Moore, Daniel Rothschild, Bernhard Salow, Miriam Schoenfield, Ginger Schultheis, James Shaw, and audiences at FEW 2020, MIT, and the Universities of Bristol, Pittsburgh, Oxford, and Sydney, for much helpful feedback and discussion.

Appendix

A.1 Deriving Deference

Recall that $q_1, ..., q_n$ are the claims that Calvin assign 80%-confidence to, that R is the rational probability function for him to have overall, and that $\overline{R}$ is the average rational confidence in the q_i: $\overline{R} := \sum_{i=1}^{n} \frac{R(q_i)}{n}$. Recall Deference:

> **Deference:** Upon learning only that the average rational confidence for Calvin to have in his 80%-opinions is x%, become x% confident in each of them.
> For all q_i: $P(q_i|\overline{R} = x) = x$.

(For simplicity of notation I maintain focus on Calvin's 80%-opinions. Obviously, parallel principles and reasoning apply to the others thresholds.)

Deference follows from two further principles:

> **Point-wise Deference:** Upon learning the rational credence function for Calvin is δ, become $\delta(q_i)$-confident in each q_i.
> For all q_i : $P(q_i|R = \delta) = \delta(q_i)$.[42]
>
> **Equality:** Upon learning only that the average rational confidence for Calvin to have in his 80% opinions is x%, be equally confident in each of them.
> For all q_i, q_j : $P(q_i|\overline{R} = x) = P(q_j|\overline{R} = x)$.

Since Equality is extremely plausible in the situations we're considering (where you don't know anything more about the q_i than they they were claims that Calvin was 80% confident in), this shows that Deference follows from the more familiar Point-wise version.

To prove this, for any random variable X (a function from possibilities to numbers), let $\mathbb{E}[X] := \sum_t P(X = t) \cdot t$ be your rational expectation of X. (Assume a finite state space, for simplicity.) Note that $\overline{R}$ is a random variable; also note that if $I(q_i)$ is the indicator variable for q_i (1 if q_i is true, 0 otherwise), then $\mathbb{E}[I(q_i)] = P(q_i)$. Let $D_x = \{\delta_1, ..., \delta_k\}$ be the set of possible values of R such that $\sum_{i=1}^{n} \frac{\delta_j(q_i)}{n} = x$, so that $\overline{R} = x \Leftrightarrow R \in D_x$.

First, focus on your expectations of the proportion of truths, conditional on $\overline{R} = x$:

$$\mathbb{E}[\sum \tfrac{I(q_i)}{n} \mid \overline{R} = x] \;=\; \sum_{\delta \in D_x} P(R = \delta | \, \overline{R} = x) \cdot \mathbb{E}[\sum \tfrac{I(q_i)}{n} \mid R = \delta]$$

[42] Here 'δ' is a rigid designator for a particular probability function (an assignment of numbers to propositions), whereas R is a definite description for "the rational credence function for Calvin, whatever it is"—so R can vary across possibilities but δ cannot.

By linearity of expectations, this equals

$$\begin{aligned}
&= \sum_{\delta \in D_x} P(R = \delta |\ \overline{R} = x) \cdot \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[I(q_i)|\ R = \delta] \\
&= \sum_{\delta \in D_x} P(R = \delta |\ \overline{R} = x) \cdot \frac{1}{n} \sum_{i=1}^{n} P(q_i|\ R = \delta) && \text{(Definition)} \\
&= \sum_{\delta \in D_x} P(R = \delta |\ \overline{R} = x) \cdot \frac{1}{n} \sum_{i=1}^{n} \delta(q_i) && \text{(Point-wise Deference)} \\
&= \sum_{\delta \in D_x} P(R = \delta |\ \overline{R} = x) \cdot x && \text{(Definition of } D_x\text{)} \\
&= x.
\end{aligned}$$

Therefore $\mathbb{E}[\sum \frac{I(q_i)}{n} \mid \overline{R} = x] = x$, so by linearity of expectations, your average rational credence in the q_i equals x: $\frac{1}{n}\sum_{i=1}^{n} P(q_i|\overline{R} = x) = x$. By Equality, since each of the values in this sum is equal, they must all be equal to x—therefore for all q_i: $P(q_i|\overline{R} = x) = x$, establishing Deference.

A.2 The Rational-to-Right Formula

Here I show how to calculate what your posterior confidence should be that Calvin is overconfident in his 80%-opinions when Deference and Independence hold, you know that there are n such opinions, and you learn how (mis)calibrated they are. Recall:

Deference: For all q_i: $P(q_i|\overline{R} = x) = x$.

Independence: For all $q_{i_0}, ..., q_{i_k}$: $P(q_{i_0}|\overline{R} = x, q_{i_1}, ..., q_{i_l}, \neg q_{i_{l+1}}, ..., \neg q_{i_k}) = P(q_{i_0}|\overline{R} = x)$

Suppose you initially leave open that $\overline{R}$ will be any of $t_1, ..., t_m$, with prior probabilities $P(\overline{R} = t_i)$. Note that Deference and Independence imply that $P(\cdot|\overline{R} = t_i)$ treats the q_i as independent, identically-distributed Bernoulli variables with success probability t_i. Letting $\overline{q}$ be the proportion of q_i that are true, that means that conditional on $\overline{R} = t_i$, $\overline{q}$ is distributed according to a binomial distribution with parameters t_i and n; in particular: $P(\overline{q} = sn|\overline{R} = t_i) = \binom{n}{sn} t_i^{sn}(1 - t_i)^{n-sn}$.

Now suppose you learn that proportion $s \cdot n$ of the q_i were true. By Bayes formula, your posterior confidence in any $\overline{R} = t_i$ hypothesis should be:

$$\begin{aligned}
P(\overline{R} = t_i|\overline{q} = sn) &= \frac{P(\overline{R} = t_i) \cdot P(\overline{q} = sn|\overline{R} = t_i)}{\sum_{j=1}^{m} P(\overline{R} = t_j) \cdot P(\overline{q} = sn|\overline{R} = t_j)} \\
&= \frac{P(\overline{R} = t_i) \cdot \binom{n}{sn} t_i^{sn}(1 - t_i)^{n-sn}}{\sum_{j=1}^{m} P(\overline{R} = t_j) \cdot \binom{n}{sn} t_j^{sn}(1 - t_j)^{n-sn}}
\end{aligned}$$

References

Anderson, John R, 1990. *The Adaptive Character of Thought.* Erlbaum Associates.

Ariely, Dan, 2008. *Predictably irrational.* Harper Audio.

Belot, Gordon, 2013a. 'Bayesian Orgulity'. *Philosophy of Science*, 80(4):483–503.

———, 2013b. 'Failure of calibration is typical'. *Statistics and Probability Letters*, 83(10):2316–2318.

Brenner, L. A., Koehler, D.J., Liberman, V., and Tversky, A., 1996. 'Overconfidence in Probability and Frequency Judgments: A Critical Examination'. *Organizational Behavior and Human Decision Processes*, 65(3):212–219.

Brenner, Lyle, 2000. 'Should Observed Overconfidence Be Dismissed as a Statistical Artifact? Critique of Erev , Wallsten , and Budescu (1994)'. 107(4):943–946.

Brenner, Lyle, Griffin, Dale, and Koehler, Derek J, 2005. 'Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment'. *Organizational Behavior and Human Decision Processes*, 97(1):64–81.

Brenner, Lyle A, Griffin, Dale W, and Koehler, Derek J, 2012. 'A Case-Based Model of Probability and Pricing Judgments : Biases in Buying and Selling Uncertainty'. 58(1):159–178.

Briggs, R., 2009a. 'Distorted Reflection'. *Philosophical Review*, 118(1):59–85.

Briggs, Ray, 2009b. 'The Anatomy of the Big Bad Bug'. *Nous*, 43(3):428–449.

Budescu, David V, Wallsten, Thomas S, and Au, Wing Tung, 1997. 'On the importance of random error in the study of probability judgment. Part II: Applying the stochastic judgment model to detect systematic trends'. *Journal of Behavioral Decision Making*, 10(3):173–188.

Carr, Jennifer Rose, 2019. 'Imprecise Evidence without Imprecise Credences'. *Philosophical Studies*, To appear.

Christensen, David, 2010a. 'Higher-Order Evidence'. *Philosophy and Phenomenological Research*, 81(1):185–215.

———, 2010b. 'Rational Reflection'. *Philosophical Perspectives*, 24:121–140.

———, 2016. 'Disagreement, Drugs, etc.: From Accuracy to Akrasia'. *Episteme*, 13(4):397–422.

Cohen, L. Jonathan, 1981. 'Can human irrationality be experimentally demonstrated?' *Behavioral and Brain Sciences*, 4(3):317–331.

Comesaña, Juan, 2020. *Being Rational and Being Right.* Oxford University Press.

Crupi, Vincenzo, Fitelson, Branden, and Tentori, Katya, 2008. 'Probability, confirmation, and the conjunction fallacy'. *Thinking & Reasoning*, 14(2):182–199.

Cushman, Fiery, 2018. 'Rationalization is rational'. 1–27.

Dawid, A P, 1982. 'The Well-Calibrated Bayesian'. *Journal of the American Statistical Association*, 77(379):605–610.

Dawid, A. P., 1983. 'Calibration-Based Empirical Inquiry'. *The Annals of Statistics*, 13(4):1251–1273.

Doody, Ryan, 2020. 'The Sunk Cost Fallacy Is Not a Fallacy'. *Ergo*, 6(40):1153–1190.

Dorst, Kevin, 2019a. 'Abominable KK failures'. *Mind*, 128(512):1227–1259.

———, 2019b. 'Evidence: A Guide for the Uncertain'. *Philosophy and Phenomenological Research*, To appear.

———, 2020. 'Higher-Order Evidence'. In Maria Lasonen-Aarnio and Clayton Littlejohn, eds., *The Routledge Handbook for the Philosophy of Evidence.* Routledge.

Dunn, Jeff, 2015. 'Reliability for degrees of belief'. *Philosophical Studies*, 172(7):1929–1952.

Dunning, David, 2012. *Self-insight: Roadblocks and detours on the path to knowing thyself.* Psychology Press.

Dunning, David, Griffin, Dale W., Milojkovic, James D, and Ross, Lee, 1990. 'The Overconfidence Effect in Social Prediction'. *Journal of Personality and Social Psychology*, 58(4):568–581.

Ehrlinger, Joyce, Mitchum, Ainsley L., and Dweck, Carol S., 2016. 'Understanding overconfidence: Theories of intelligence, preferential attention, and distorted self-assessment'. *Journal of Experimental Social Psychology*, 63:94–100.

Elga, Adam, 2013. 'The puzzle of the unmarked clock and the new rational reflection principle'. *Philosophical Studies*, 164(1):127–139.

———, 2016. 'Bayesian Humility'. *Philosophy of Science*, 83(3):305–323.

Erev, Ido, Wallsten, Thomas S, and Budescu, David V, 1994. 'Simultaneous over-and underconfidence: The role of error in judgment processes.' *Psychological review*, 101(3):519.

Fine, Cordelia, 2005. *A Mind of its Own: How Your Brain Distorts and Deceives.* W. W. Norton & Company.

Fitelson, Branden and Hawthorne, James, 2010. 'The Wason Task(s) and the Paradox of Confirmation'. *Philosophical Perspectives*, 24:207–241.

Gallow, J. Dmitri, 2019. 'Updating for externalists'. *Noûs*, (November 2018):1–30.

Gigerenzer, Gerd, 1991. 'How to make cognitive illusions disappear: Beyond heuristics and biases'. *European review of social psychology*, 2(1):83–115.

Gigerenzer, Gerd, Hoffrage, Ulrich, and Kleinbölting, Heinz, 1991. 'Probabilistic mental models: a Brunswikian theory of confidence.' *Psychological review*, 98(4):506.

Glaser, Markus and Weber, Martin, 2007. 'Overconfidence and trading volume'. *The Geneva Risk and Insurance Review*, 32(1):1–36.

———, 2010. 'Overconfidence'. *Behavioral finance: Investors, corporations, and markets*, 241–258.

Greco, Daniel and Hedden, Brian, 2016. 'Uniqueness and metaepistemology'. *The Journal of Philosophy*, 113(8):365–395.

Griffin, Dale and Tversky, Amos, 1992. 'The Weighing of Evidence and the Determinants of Confidence'. *Cognitive Psychology*, 24:411–435.

Griffiths, Thomas L., Chater, Nick, Norris, Dennis, and Pouget, Alexandre, 2012. 'How the bayesians got their beliefs (and what those beliefs actually are): Comment on bowers and davis (2012)'. *Psychological Bulletin*, 138(3):415–422.

Hahn, Ulrike and Harris, Adam J.L., 2014. 'What Does It Mean to be Biased. Motivated Reasoning and Rationality.' In *Psychology of Learning and Motivation - Advances in Research and Theory*, volume 61, 41–102.

Hahn, Ulrike and Oaksford, Mike, 2007. 'The rationality of informal argumentation: a Bayesian approach to reasoning fallacies.' *Psychological review*, 114(3):704.

Hall, Ned, 1994. 'Correcting the Guide to Objective Chance'. *Mind*, 103(412):505–517.

Harris, Adam J L and Hahn, Ulrike, 2011. 'Unrealistic optimism about future life events: A cautionary note.' *Psychological review*, 118(1):135.

Harvey, Nigel, 1997. 'Confidence in judgment'. *Trends in cognitive sciences*, 1(2):78–82.

Hastie, Reid and Dawes, Robyn M, 2009. *Rational choice in an uncertain world: The psychology of judgment and decision making*. Sage Publications.

Hedden, Brian, 2018. 'Hindsight Bias is not a Bias'. *Analysis*, To appear.

Hoffrage, Ulrich, 2004. 'Overconfidence'. In *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*, 235–254.

Horowitz, Sophie, 2014a. 'Epistemic Akrasia'. *Noûs*, 48(4):718–744.

———, 2014b. 'Immoderately rational'. *Philosophical Studies*, 167:41–56.

———, 2019a. 'Predictably Misleading Evidence'. In Mattias Skipper and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidence: New Essays*, 105–123. Oxford University Press.

———, 2019b. 'The Truth Problem for Permissivism'. *The Journal of Philosophy*, 116(5):237–262.

Howard, Michael, 1984. *The Causes of Wars and Other Essays*. Harvard University Press.

Icard, Thomas, 2017. 'Bayes, Bounds, and Rational Analysis'. *Philosophy of Science*, 694837.

Isaacs, Yoaav, 2019. 'The Fallacy of Calibrationism'. *Philosophy and Phenomenological Research*, To appear.

Jansen, Rachel, Rafferty, Anna N, and Griffiths, Tom, 2018. 'Modeling the Dunning-Kruger Effect: A Rational Account of Inaccurate Self-Assessment.' In *CogSci*.

Johnson, Dominic D P, 2009. *Overconfidence and war*. Harvard University Press.

Johnson, Dominic D.P. and Fowler, James H., 2011. 'The evolution of overconfidence'. *Nature*, 477(7364):317–320.

Joyce, James M, 1998. 'A Nonpragmatic Vindication of Probabilism'. *Philosophy of Science*, 65(4):575–603.

Juslin, Peter, 1994. 'The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items'. *Organizational Behavior and Human Decision Processes*, 57(2):226–246.

Juslin, Peter, Olsson, Henrik, and Björkman, Mats, 1997. 'Brunswikian and thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgment'. *Journal of Behavioral Decision Making*, 10(3):189–209.

Juslin, Peter, Wennerholm, Pia, and Olsson, Henrik, 1999. 'Format Dependence in Subjective Probability Calibration'. *Journal of Experimental Psychology: Learning Memory and Cognition*, 25(4):1038–1052.

Juslin, Peter, Winman, Anders, and Olsson, Henrik, 2000. 'Naive empiricism and dogmatism in confidence research: A critical examination of the hardeasy effect.' *Psychological review*, 107(2):384.

Kahneman, Daniel, 2011a. 'Don't Blink! The Hazards of Confidence'.

———, 2011b. *Thinking Fast and Slow*. Farrar, Straus, and Giroux.

Kahneman, Daniel, Slovic, Paul, and Tversky, Amos, eds., 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Kahneman, Daniel and Tversky, Amos, 1996. 'On the reality of cognitive illusions.'

Kelly, Thomas, 2004. ‘Sunk costs, rationality, and acting for the sake of the past’. *Nous*, 38(1):60–85.

———, 2008. ‘Disagreement, Dogmatism, and Belief Polarization’. *The Journal of Philosophy*, 105(10):611–633.

Keren, Gideon, 1987. ‘Facing uncertainty in the game of bridge: A calibration study’. *Organizational Behavior and Human Decision Processes*, 39(1):98–114.

Koehler, Derek J, Brenner, Lyle, and Griffin, Dale, 2002. ‘The calibration of expert judgment: Heuristics and biases beyond the laboratory’. *Heuristics and biases: The psychology of intuitive judgment*, 686–715.

Koralus, Philipp and Mascarenhas, Salvador, 2013. ‘The erotetic theory of reasoning: bridges between formal semantics and the psychology of deductive inference’. *Philosophical Perspectives*, 27:312–365.

Kruger, Justin and Dunning, David, 1999. ‘Unskilled and Unaware of it: How Difficulties in Recognizing One’s Own Incompetence Lead to Inflated Self-Assessments’. *Journal of Personality and Social Psychology*, 77(6):121–1134.

Lam, Barry, 2011. ‘On the rationality of belief-invariance in light of peer disagreement’. *Philosophical Review*, 120(2):207–245.

———, 2013. ‘Calibrated probabilities and the epistemology of disagreement’. *Synthese*, 190(6):1079–1098.

Lasonen-Aarnio, Maria, 2013. ‘Disagreement and evidential attenuation’. *Nous*, 47(4):767–794.

———, 2015. ‘New Rational Reflection and Internalism about Rationality’. In Tamar Szabó Gendler and John Hawthorne, eds., *Oxford Studies in Epistemology*, volume 5, 145–171. Oxford University Press.

———, 2019. ‘Higher-Order Defeat and Evincibility’. *Higher-Order Evidence: New Essays*, 144.

Lewis, David, 1980. ‘A subjectivist’s guide to objective chance’. In Richard C Jeffrey, ed., *Studies in Inductive Logic and Probability*, volume 2. University of California Press.

———, 1994. ‘Humean Supervenience Debugged’. *Mind*, 103(412):473–490.

Lewis, Michael, 2016. *The undoing project: A friendship that changed the world*. Penguin UK.

Lichtenstein, Sarah, Fischhoff, Baruch, and Phillips, Lawrence D., 1982. ‘Calibration of probabilities: The state of the art to 1980’. In Daniel Kahneman, Paul Slovic, and Amos Tversky, eds., *Judgment under Uncertainty*, 306–334. Cambridge University Press.

Littlejohn, Clayton, 2012. *Justification and the truth-connection*. Cambridge University Press.

Magnus, Jan R. and Peresetsky, Anatoly A., 2018. ‘Grade expectations: Rationality and overconfidence’. *Frontiers in Psychology*, 8(JAN):1–10.

Mahtani, Anna, 2017. ‘Deference, respect and intensionality’. *Philosophical Studies*, 174(1):163–183.

Mandelbaum, Eric, 2018. ‘Troubles with Bayesianism: An introduction to the psychological immune system’. *Mind & Language*, 1–17.

Merkle, Christoph and Weber, Martin, 2011. ‘True overconfidence: The inability of rational information processing to account for apparent overconfidence’. *Organizational Behavior and Human Decision Processes*, 116(2):262–271.

Moore, Don A and Healy, Paul J, 2008. ‘The trouble with overconfidence.’ *Psychological review*, 115(2):502.

Moore, Don A, Tenney, Elizabeth R, and Haran, Uriel, 2015. ‘Overprecision in judgment’. *The Wiley Blackwell handbook of judgment and decision making*, 2:182–209.

Myers, David G., 2010. *Psychology*. Worth Publishers, ninth edit edition.

Nebel, Jacob M., 2015. ‘Status quo bias, rationality, and conservatism about value’. *Ethics*, 125(2):449–476.

Oaksford, Mike and Chater, Nick, 1994. ‘A Rational Analysis of the Selection Task as Optimal Data Selection’. *Psychological Review*, 101(4):608–631.

———, 2007. *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

O’Connor, Cailin and Weatherall, James Owen, 2018. ‘Scientific Polarization’. *European Journal for Philosophy of Science*, 8(3):855–875.

Odean, Terrance, 1999. ‘Do Investors Trade Too Much?’ *American Economic Review*, 89(5):1279–1298.

Ortoleva, Pietro and Snowberg, Erik, 2015. ‘Overconfidence in political behavior’. *American Economic Review*, 105(2):504–535.

Pettigrew, Richard, 2016a. *Accuracy and the Laws of Credence*. Oxford University Press.

———, 2016b. ‘Jamesian Epistemology Formalized: An Explication of ‘The Will to Believe’’. *Episteme*, 13(3):253–268.

Pettigrew, Richard and Titelbaum, Michael G, 2014. ‘Deference Done Right’. *Philosopher’s Imprint*, 14(35):1–19.

Pfeifer, Phillip E, 1994. 'Are we overconfident in the belief that probability forecasters are overconfident?' *Organizational Behavior and Human Decision Processes*, 58(2):203–213.

Plous, Scott, 1993. *The psychology of judgment and decision making.* Mcgraw-Hill Book Company.

Quilty-Dunn, Jake, 2020. 'Unconscious Rationalization, or: How (Not) To Think About Awfulness and Death'.

Roush, Sherrilyn, 2009. 'Second Guessing: A Self-Help Manual'. *Episteme*, 251–268.

———, 2016. 'Knowledge of Our Own Beliefs'. *Philosophy and Phenomenological Research*, 93(3).

———, 2017. 'Epistemic Self-Doubt'.

Salow, Bernhard, 2018. 'The Externalist's Guide to Fishing for Compliments'. *Mind*, 127(507):691–728.

Schoenfield, Miriam, 2012. 'Chilling out on epistemic rationality'. *Philosophical Studies*, 158(2).

———, 2014. 'Permission to Believe: Why Permissivism is True and What it Tells Us About Irrelevant Influences on Belief'. *Nous*, 48(2):193–218.

———, 2015. 'A Dilemma for Calibrationism'. *Philosophy and Phenomenological Research*, 91(2):425–455.

———, 2016a. 'An Accuracy Based Approach to Higher Order Evidence'. *Philosophy and Phenomenological Research*, To Appear.

———, 2016b. 'Bridging Rationality and Accuracy'. *Journal of Philosophy*, 112(12):633–657.

———, 2019. 'Permissivism and the value of rationality: A challenge to the uniqueness thesis'. *Philosophy and phenomenological research*, 99(2):286–297.

Schultheis, Ginger, 2018. 'Living on the Edge: Against Epistemic Permissivism'. *Mind*, 127(507):863–879.

Seidenfeld, Teddy, 1985. 'Calibration , Coherence , and Scoring Rules'. *Philosophy of Science*, 52:274–294.

Shariatmadari, David, 2015. 'Daniel Kahneman: What would I eliminate if I had a magic wand? Overconfidence''.

Singer, Daniel J, Bramson, Aaron, Grim, Patrick, Holman, Bennett, Jung, Jiin, Kovaka, Karen, Ranginani, Anika, and Berger, William J, 2019. 'Rational social and political polarization'. *Philosophical Studies*, 176(9):2243–2267.

Sliwa, Paulina and Horowitz, Sophie, 2015. 'Respecting *all* the evidence'. *Philosophical Studies*, 172(11):2835–2858.

Staffel, Julia, 2020. *Unsettled thoughts: A theory of degrees of rationality.* Oxford University Press, USA.

Stich, Stephen P., 1985. 'Could Man be an Irrational Animal?' *Synthese*, 64:115–135.

Tenenbaum, Joshua B and Griffiths, Thomas L, 2006. 'Optimal Predictions in Everyday Cognition'. *Psychological Science*, 17(9):767–773.

Tenenbaum, Joshua B, Kemp, Charles, Griffiths, Thomas L, and Goodman, Noah D, 2011. 'How to grow a mind: Statistics, structure, and abstraction'. *science*, 331(6022):1279–1285.

Tetlock, Philip E and Gardner, Dan, 2016. *Superforecasting: The art and science of prediction.* Random House.

Thaler, Richard H., 2015. *Misbehaving: The Making of Behavioural Economics.* Penguin.

Tversky, Amos and Kahneman, Daniel, 1974. 'Judgment under uncertainty: Heuristics and biases'. *Science*, 185(4157):1124–1131.

———, 1983. 'Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment.' *Psychological review*, 90(4):293.

Vallone, Robert P., Griffin, Dale W., Lin, Sabrina, and Ross, Lee, 1990. 'Overconfident Prediction of Future Actions and Outcomes by Self and Others'. *Journal of Personality and Social Psychology*, 58(4):582–592.

van Fraassen, Bas, 1983. 'Calibration: A Frequency Justification for Personal Probability'. In R.S. Cohen and L Laudan, eds., *Physics, Philosophy, and Psychoanalysis: Essays in Honor of Adolf Gr unbaum*, 295–318. D. Reidel Publishing Company.

———, 1984. 'Belief and the Will'. *The Journal of Philosophy*, 81(5):235–256.

van Prooijen, Jan-Willem and Krouwel, André P M, 2019. 'Psychological Features of Extreme Political Ideologies'. *Current Directions in Psychological Science*, 28(2):159–163.

White, Roger, 2005. 'Epistemic Permissiveness'. *Philosophical Perspectives*, 19(1):445–459.

———, 2009a. 'Evidential Symmetry and mushy credence'. *Oxford Studies in Epistemology*, 161–186.

———, 2009b. 'On Treating Oneself and Others as Thermometers'. *Episteme*, 6(3):233–250.

Williamson, Timothy, 2000. *Knowledge and its Limits.* Oxford University Press.

———, 2019. 'Evidence of Evidence in Epistemic Logic'. In Mattias Skipper and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidence: New Essays*, 265–297. Oxford University Press.