Higher-Order Evidence

Kevin Dorst
kevindorst@pitt.edu

July 2020

Forthcoming in Maria Lasonen-Aarnio and Clayton Littlejohn (eds.),
The Routledge Handbook for the Philosophy of Evidence, Routledge.

Abstract

On at least one of its uses, 'higher-order evidence' refers to evidence about what opinions are rationalized by your evidence. This chapter surveys the foundational epistemological questions raised by such evidence, the methods that have proven useful for answering them, and the potential consequences and applications of such answers.

Higher-order evidence is evidence about evidence. But although that may seem clear enough, the term has been used to refer to (at least) two different things:

(1) Evidence about whether various responses to your evidence are *reliable*.

(2) Evidence about whether various responses to your evidence are *rational*.

These types of evidence will *often* go together: since rational opinions are correlated with truth, evidence about rationality is usually evidence about reliability. But they can clearly come apart. *Case 1:* you pick up a textbook and start believing its claims—yet unbeknownst to you, it's full of fabrications. Then your resulting opinions are (1) systematically *un*reliable but (2) perfectly rational. *Case 2:* an apparent crackpot emails you his magnum opus predicting the future, and you blithely believe his every word—yet, contrary to all your evidence, it turns out he *can* see the future. Then your resulting opinions are (1) systematically reliable but (2) completely *ir*rational. Thus evidence that you're in Case 1 is "higher-order evidence" that your opinions were unreliable but rational; evidence that you're in Case 2 is "higher-order evidence" that your opinions were reliable but irrational.

Upshot: it's important to keep these two notions of "higher-order evidence" conceptually distinct. Some discussions have focused on (1)[1], while others have focused

[1]E.g. Elga (2007); Christensen (2007c, 2010a); Lasonen-Aarnio (2010); Christensen (2016); Roush (2009, 2016, 2017); White (2009b); Schoenfield (2015, 2016a); Sliwa and Horowitz (2015); Isaacs (2019).

on (2).[2] This chapter will offer an opinionated summary of recent work on (2): How do rational people respond to evidence about what opinions are rationalized by their evidence? We'll cover the basic questions that arise (§1), a fruitful set of methods for making progress on them (§2), a map of the current theoretical positions (§3), and a sketch of the relevance of these positions to other debates (§4).

1 The Questions

The higher-order evidence literature is often motivated by a dilemma that takes (something like) the following form.[3] Rationally using our evidence is hard, so we should often be unsure whether we've messed it up. As a result, we can receive evidence that we *have* messed it up—even if, in fact, we have not. For example, suppose you just completed a math problem and you're fairly confident in your answer of '427'—but you're also sleep-deprived, and so are unsure if that confidence is rationally based on the evidence. Suppose that in fact it is, but then your TA (mistakenly) tells you that your answer is way off. This provides you with (misleading) higher-order evidence that your evidence *didn't* rationalize confidence in '427'. More generally:

(i) It's possible to get misleading higher-order evidence: to have evidence that supports q, but then get evidence that makes it rational to believe that one's evidence didn't support q.

This higher-order evidence seems like it could make it rational to believe "My evidence didn't support '427'." Nevertheless, your original (so-called "first-order") evidence is still there—for example, the parameters of the math problem. Thus it seems that your total evidence might still support the claim that '427' is the right answer; after all, if it's a math problem, the evidence seems to entail as much! Putting these pieces together, it seems you could have total evidence supporting "'427' is the right answer, but my evidence doesn't support that." More generally:

(ii) From (i) it follows that it's possible to have evidence that rationalizes believing "q but my evidence doesn't support q".[4]

Yet believing "'427' is the answer but my evidence doesn't support that" seems like it would lead to clearly irrational behavior and beliefs—such as betting that your own evidence is misleading (Horowitz 2014). That supports:

[2]E.g. Williamson (2000, 2008, 2014, 2019); Briggs (2009); Christensen (2010b); Smithies (2012, 2015); Lasonen-Aarnio (2013, 2014, 2015, 2019); Greco (2014); Horowitz (2014, 2019); Pettigrew and Titelbaum (2014); Titelbaum (2015); Schoenfield (2016b, 2017a); Littlejohn (2018); Schultheis (2018); Salow (2018b, 2019); Skipper et al. (2018); Carr (2019a,b); Eder and Brössel (2019); Gallow (2019a,b); Das (2020a,b); Fraser (2020).

[3]E.g. Greco (2014, 2019); Neta (2019); Littlejohn (2018); Worsnip (2018); Kappel (2019); Salow (2019); Skipper (2019b); Lasonen-Aarnio (2020).

[4]See Worsnip (2018, 2019); Skipper (2019a); Lasonen-Aarnio (2020) for more on (ii).

(iii) But the "epistemically akratic" state in (ii) *can't* be rational.

And thus we have a dilemma.

The literature contains a multitude of responses. Rather than try to survey them, I'll explain why the underlying issue—the plausibility of (i)—opens up a plethora of interesting questions in traditional, formal, and applied epistemology.

The issue I have in mind arises within a *single* normative notion, rather than as a tension between differing normative notions. So pick your favorite: "what's supported by the evidence"; "what you should think"; "what would be rational to think"; etc. I'll use these interchangeably.

The issue is: What is the import of **higher-order uncertainty**? First-order uncertainty is the kind much of epistemology focuses on: Will it rain tomorrow? Who will win the next presidential election? Do my colleagues think I smell? Etc. *Higher-order* uncertainty is uncertainty about what degree of uncertainty in these claims is rational: Should I be 50-50 or 60-40 that it'll rain tomorrow? Should I be confident it'll be a Democrat, or am I biased in my judgment? Should I wear more deodorant—or am I paranoid in my suspicions about my colleagues? Etc. Your evidence rationalizes higher-order uncertainty when it rationalizes being unsure what opinions your evidence rationalizes—or, more colloquially, when you should be unsure what you should think.[5]

How do we think precisely about higher-order uncertainty (and, thereby, higher-order evidence)? Higher-order uncertainty arises for a given normative notion when it fails to be *introspective*: when there is a rational (evidentially supported, etc.) failure to be certain of what is rational (evidentially supported, etc.). This can happen for any representation—outright belief, comparative confidence, mushy credence, etc.—but thus far the most fruitful way of modeling higher-order uncertainty has emerged within a framework that uses a unique (White 2005), precise (White 2009a) probability function P that encodes the rational degrees of belief.

'P' is a definite description for 'the rational credence function, whatever it is (given your evidence, now)'. So '$P(q) = 0.6$' is true iff it's rational to be 0.6 confident in q given your evidence. Likewise, '$P(P(q) = 0.6) = 0.2$' is true iff it's rational to be 0.2 confident that it's rational to be 0.6 confident in q, and so on. Thus it's rational to have higher-order uncertainty iff there is a proposition q and a threshold t such that you should be unsure whether you should be t-confident of q: $0 < P(P(q) = t) < 1$.[6]

First question: can higher-order uncertainty be rational? If we say *no*, then we have

[5]Why have we moved from talk of higher-order *evidence* to higher-order *uncertainty*? Because you can get (nontrivial) evidence about X iff you can be uncertain about X. This falls out of a Bayesian picture of evidence, but is also independently motivated: evidence is something that can *change your mind*—and if you can't be uncertain about X, you can't change your mind about X.

[6]Don't confuse P with your *actual* credences C: you can perfectly well be rational ($P = C$), know what your credences are ($C(q) = t \rightarrow C(C(q) = t) = 1$), and be unsure whether those credences are rational ($C(P = C) < 1$); see Dorst (2019b).

adopted a **Merging** theory, collapsing all levels of uncertainty down to the first one. Our formal theory of higher-order uncertainty is simple:

> **Access Internalism:** If $P(q) = t$, then $P(P(q) = t) = 1$.
> If you should be t-confident of q, you should be certain that you should be.

If this is our theory, our solution to the dilemma posed above is simply to deny (i): it turns out that it's really not possible to receive misleading higher-order evidence after all. The main challenge for this approach is to explain *why* higher-order uncertainty is irrational, and to build a philosophical theory that explains our apparently rational sensitivity to it—as when, for instance, we are sensitive to the disagreement of peers who share our evidence.[7]

Suppose instead we agree that higher-order uncertainty is sometimes rational—and therefore that Access Internalism is false. Then the pressing questions is: *what kinds* of higher-order uncertainty can be rational? The opposite extreme of Merging is a **Splitting** theory, which says that necessary connections between first and higher-order opinions are either minimal or nonexistent.[8] Splitting theories deny (iii) in our above dilemma, allowing that it can sometimes be rational to be epistemically akratic. More generally, they allow that each of the following can sometimes be rational (see §2):

> **Living on the Edge:** Having a given credence in q while being certain that the rational credence is no higher, and leaving open that it's lower.
> (There are q and t, such that $P(q) = t$ but $P(P(q) \leq t) = 1$ and $P(P(q) < t) > 0$.)[9]

> **Unconfident Confidence:** Having confidence in q without being confident that it's rational to have confidence in q.
> (There are q and t such that $P(q) \geq t$ but $P(P(q) \geq t) < t$.)[10]

> **Improbable Certainties:** Being certain of q but having some credence that you shouldn't be.
> (There are q and $t \geq 0$ such that $P(q) = 1$ but $P(P(q) < 1) > t$.)[11]

[7] Examples of Mergers: Smithies (2012, 2015, 2019); Greco (2014); Titelbaum (2015, 2019); Schoenfield (2017a); Salow (2018b, 2019); Tal (2018, 2020); Skipper (2020). See Williamson (2008); Lasonen-Aarnio (2013); Roush (2016); Dorst (2019b, 2020) for some of the challenges.

[8] A standard Splitting position: the only connection is that if your evidence makes q certain, then q is true. The probabilistic component of this is the "surely-factivity" axiom: $P(q|P(q) = 1) = 1$. See Williamson (2000, 2008, 2014, 2019); Lasonen-Aarnio (2013, 2015); Weatherson (2019).

[9] Similar principles are discussed in Christensen (2007a, 2010b); Sliwa and Horowitz (2015); Schultheis (2018); Skipper et al. (2018); Hawthorne and Isaacs (2020); Fraser (2020); Dorst (2021).

[10] Equivalently, there are r and s such that $P(P(r) > s) > s$ but $P(r) \leq s$. (Let $r = \neg q$, $s = 1 - t$: $[P(q) \geq t] \wedge [P(P(q) \geq t) < t]$ iff $[P(\neg q) \leq 1 - t] \wedge [P(P(q) < t) > 1 - t]$, iff $[P(r) \leq s] \wedge [P(P(r) > s) > s]$.) See (Dorst 2020, Fact 5.5); Littlejohn (2019); Littlejohn and Dutant (2019); Williamson (2019).

[11] The phenomenon (with $t \geq \frac{1}{2}$) is discussed (with a focus on knowledge rather than rational certainty) in Williamson (2013, 2014); Cohen and Comesaña (2013); Goodman (2013).

Confidence Akrasia: Being confident of q *but I shouldn't be confident of* q. (There are q and t such that $P(q \wedge [P(q) < t]) \geq t$.)[12]

Effacing Evidence: Being sure that your evidence is misleading about q. (There are q and t such that $P(q \leftrightarrow [P(q) < t]) = 1$ and $P(\neg q \leftrightarrow [P(q) > t]) = 1$.)[13]

These "mismatches" between first- and higher-order opinions can lead to surprisingly counterintuitive effects. For instance, if you have Effacing Evidence, then receiving higher-order evidence that your evidence supports confidence in q (i.e. that $P(q) > t$) should lead you to *lower* your confidence in q.

We might be inclined to deny that some of these are genuinely possible. If so, then we've moved from Splitting to **Bridging**: a theory that asserts that although higher-order uncertainty can be rational, there are still systematic connections between first- and higher-order opinions. Though intuitive, Bridging faces the challenge of drawing and defending a principled line that explains why some types of higher-order uncertainty can be rational, and others cannot. Much of what follows will focus on delineating the possible such lines.

Summing up: the puzzle of higher-order evidence raises the question of which sorts of higher-order uncertainty can be rational. The more higher-order uncertainty that a theory allows, the weaker the connections it imposes between higher- and first-order evidence; the less higher-order uncertainty it allows, the stronger those connections. The role of a theory is to explain which types of higher-order uncertainty (hence higher-order evidence) are possible, and why.

Next question: How can we decide between such theories?

2 The Methods

The literature has used a mix of different methods, moving between (a) formulating and testing general principles, (b) examining simple cases, and (c) building toy formal models of those simple cases. The most fruitful approach involves using these methods in tandem: general principles are formulated to delineate and constrain our theory; cases are offered to stress test those principles; toy models are constructed to check the coherence of those cases and the tenability of those principles.

It may be obvious why we need both (a) general principles and (b) cases—but why do we need (c) models? The relevant sense of 'model' is a specific one familiar from logic—for example, possible-worlds models from modal logic. These are formal constructs that

[12] This is effectively the denial of (iii) from our dilemma, though it involves akratic *confidence* (Horowitz 2014); variants involve akratic beliefs (Greco 2014) or estimates (Christensen 2010b; Salow 2019; Dorst 2019b). Note that the denial of Confidence Akrasia for all q and t is equivalent to the assertion of $P([P(q) \geq t] \rightarrow q) \geq t$, which is a material-conditional version of the "Trust" principle below (§3).

[13] See Horowitz (2014); Ahmed and Salow (2018); Salow (2018b); Dorst (2020).

are used to interpret a (formal or informal) language, and thereby used to check whether claims in that language are consistent, true, etc. We need them because we often want to know whether various described situations are consistent with background principles, and yet it is often hard to tell. For example, is the situation in (1) consistent with the principle in (2)?

(1) You are rational to be unsure whether the rational credence function is your own, or is one that assigns a credence to q that is a bit higher or a bit lower.

(2) You should always match your credence in a claim to your best estimate of the rational credence in that claim. (See the 'Estimate-Enkrasia' principle in §3.)

They certainly *look* consistent.[14] After all, you might have a credence of 0.5 in q, and be equally unsure whether the rational credence in q is 0.4, 0.5, or 0.6. Then your credence in q equals your best estimate of the rational credence in q. But this only checks one instance of (2), which is a universally quantified claim. And it turns out that if we try to draw a full model that makes both true, and we'll find that we cannot—surprisingly, (1) and (2) are *in*consistent, for (2) implies that higher-order uncertainty is irrational (Samet 2000; Dorst 2019b).

Upshot: we can't assess the plausibility of principles about higher-order uncertainty by simply trying out a few of their instances; we need to build models of them so that we can systematically check whether *all* their instances are satisfiable.

What do such models look like? The conceptually simplest ones are **probability frames** $\langle W, P\rangle$, which consist of a set of worlds W and a function P from worlds w to probability distributions P_w over W.[15] Let's assume that only finitely many distinctions are relevant, so that W is finite and we can list all the worlds in some order $w_1, ..., w_n$. Propositions are modeled as sets of worlds, with logical operations done via set theory: q is true at w iff $w \in q$; $\neg q := W - q$; $p \wedge q := p \cap q$; $p \rightarrow q := \neg p \cup q$; etc.

The intended interpretation of P is that P_w is the rational credence function to have at w, so that 'P' is a definite description—'the rational credence function, whatever it is'. Since W is finite, P_w can be thought of as simply an assignment of non-negative numbers to the w_i that sum to 1, and for any proposition (set of worlds) q, $P_w(q) := \sum_{w_i \in q} P_w(w_i)$. P can then be used—in the way that is standard from modal logic—to define propositions about probabilities as sets of worlds in the frame. For any $q \subseteq W$ and number t, $[\boldsymbol{P(q) = t}] := \{w \in W | P_w(q) = t\}$.[16] (Note: letting wRx

[14] And many seem to have presupposed that they are: (Christensen 2007b, 2010b; Pettigrew and Titelbaum 2014; Sliwa and Horowitz 2015; Roush 2016; Skipper et al. 2018; Gallow 2019a).

[15] Dorst (2019b) gives an accessible introduction to these models; see also Gaifman (1988); Samet (2000); Dorst (2020); Levinstein (2019). Note: in the modal-logic jargon, these are "frames", not "models", because they do not include an interpretation function. As we'll see, it's not necessary.

[16] As a formal language would get cumbersome here, we use a semi-formal one. 'P' is a definite description for 'the rational credence function, whatever it is'; 'P_w' rigidly picks out the rational credence function associated with world w; for any $q \subseteq W$, '$P(q)$' is a definite description for 'the rational credence

hold iff $P_w(x) > 0$, a probability frame induces a standard Kripke frame $\langle W, R\rangle$, and $[P(\cdot) = 1]$ becomes a necessity operator with the standard Kripke semantics: $[P(q) = 1]$ is true at w iff $\forall x : \ wRx \rightarrow x \in q$.)

Here's a simple example. Suppose you're presented with an argument for a hard-to-assess claim—say, that a universal basic income is financially feasible. It's natural to think that there may be an epistemic asymmetry when it comes to evaluating such an argument. In general, one can often tell good and bad arguments apart—good arguments generally seem good; bad arguments generally seem bad. But when it comes to the sorts of arguments you're likely to come across in public discourse, there is a selection effect that distorts that tendency. Good arguments tend to be repeated because they *are* good; bad arguments tend to be repeated because they *seem* good. Thus amongst the arguments you're liable to run into, good ones will tend to wear their goodness on their sleeves, whereas bad ones will tend not to wear their badness on their sleeves—instead, they'll be masquerading as good ones (Dorst 2019c).

Here's a toy model of this epistemic asymmetry. There are two possibilities: the argument is either a good (g) or bad (b) argument. By the above asymmetry, how confident you should be in these two possibilities depends on which possibility you're in. If the argument is good, you should be (say) 80% confident that it's good, while if the argument is bad, you should be only (say) 60% confident that it's bad. This specifies a probability frame, which we can diagram as follows, drawing labeled arrows between worlds to indicate how much probability the origin world assigns to the target world:

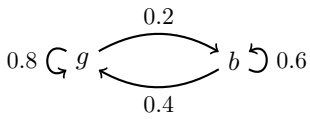

Figure 1: Good/Bad Argument

In this toy model, if the argument is good (you're at g), you should be 80% confident that the argument is good and 20% confident that it's bad: $P_g(g) = 0.8$ and $P_g(b) = 0.2$. Meanwhile, if the argument's bad (you're at b), you should be 40% confident that it's good and 60% confident that it's bad: $P_b(g) = 0.4$ and $P_b(b) = 0.6$. Thus the proposition that *you should be* 80% *confident it's good* is true only at g: $[P(g) = 0.8] = \{w|P_w(g) = 0.8\} = \{g\}$. Meanwhile, the proposition that *you should be* 40% *confident it's good* is true only at b: $[P(g) = 0.4] = \{w|P_w(g) = 0.4\} = \{b\}$. That means, in turn, that at g it's rational for you to be 80% confident that it's rational for you to be 80% confident that it's good: $P_g(P(g) = 0.8) = P_g(\{g\}) = 0.8$. Similarly, at b it's rational for you to be 40% confident that it's rational for you to be 80% confident that it's

in q, whatever it is'; '$P_w(q)$' rigidly picks out a particular number; and for any $q \subseteq W$ and $t \in [0, 1]$, '$[P(q) = t]$' picks out the set of worlds w such that $P_w(q) = t$.

good: $P_b(P(g) = 0.8) = P_b(\{g\}) = 0.4$. Likewise: if the argument is good, it's rational for you to be 20% confident that it's rational for you to be 40% confident that it's rational for you to be 80% confident that it's good: $P_g(P(P(g) = 0.8) = 0.4) = 0.2$, since $[P(P(g) = 0.8) = 0.4] = \{b\}$. (Didn't I tell you we need models?)

In short, P is used to identify propositions about probabilities as sets of worlds in the frame, which in turn allows us to "unravel" higher-order claims of any complexity to be claims about what world you're in. This is as it should be: the central insight of Hintikka and Kripke semantics is that what world you're in determines what you know (what the rational credence function is; what is necessary; etc.); therefore, claims about the latter can be reduced to claims about the former (Hintikka 1962; Kripke 1963). We are simply embedding this standard modal-logic idea in a probabilistic framework.[17] Although the mathematics may seem intimidating, it's much simpler than it looks: all that's needed is the basics of propositional modal logic and finite probability theory.[18]

A useful special case of probability frames are **prior frames**: probability frames in which the probability distributions at different worlds can each be recovered by conditioning a common prior distribution on different (propositional) pieces of evidence. Prior frames are natural for modeling situations in which you should be unsure what your evidence *is* (what you should condition on), but should be sure what various bodies of evidence *support* (what the rational prior is). Such frames can be represented with triples $\langle W, E, \pi \rangle$, where $\langle W, E \rangle$ is a standard (finite) Kripke frame in which E is a (serial[19]) binary relation between worlds. Let E_w be the set of accessible worlds from w, i.e. $\boldsymbol{E_w} := \{x \in W | wEx\}$. π is a regular[20] "prior" probability distribution over W, and the probability functions at each world are recovered by conditioning the prior on the set of accessible worlds: $P_w(\cdot) := \pi(\cdot | E_w)$. Prior frames are formally useful because it's often possible to characterize probabilistic principles within prior frames making reference only to the characteristics of the binary relation E, without reference to π.[21]

[17] This basic move for studying higher-order uncertainty is common (Aumann 1976; Gaifman 1988; Geanakoplos 1989; Samet 1999, 2000; Williamson 2000, 2008, 2014, 2019; Cresto 2012; Lasonen-Aarnio 2013, 2015; Lederman 2015; van Ditmarsch et al. 2015; Campbell-Moore 2016; Salow 2018b, 2019; Das 2020a,b; Dorst 2020, 2019b; Levinstein 2019). Notably, probability frames are formally identical to what are known as *Markov chains*—though their intended interpretation is different. As such, an n-world probability frame can be represented with an $n \times n$ *stochastic matrix* in which the rows contain non-negative entries that sum to 1. Row i, column j of the matrix represents the probability that world i assigns to world j, i.e. $P_{w_i}(w_j)$. For instance our Good/Bad Argument frame can be represented as: $\begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{pmatrix}$. This representation is useful for bookkeeping and allows for the application (e.g. Samet 2000; Williamson 2008) of some general theorems (Bertsekas and Tsitsiklis 2008, Ch. 7).

[18] The first two chapters of Bertsekas and Tsitsiklis (2008)—a textbook freely available online—covers all the relevant material on probability; no prior mathematics needed. The other half is covered by the propositional-modal-logic portion of any standard textbook (e.g. Hughes et al. 1996; Sider 2010).

[19] For all $w \in W$ there is an $x \in W$ such that wEx.

[20] For all $w \in W : \ \pi(w) > 0$.

[21] See Geanakoplos (1989); Williamson (2000, 2008, 2014, 2019); Lasonen-Aarnio (2015); Salow (2018b, 2019); Dorst (2020); Das (2020a,b).

Let's look at another example. Legend has it that Tim Williamson—a well-known trickster—owns an unmarked clock which has a single hand that can be set to any one of the twelve hour-positions. Every day he sets the position randomly, so in the morning you should be equally confident that it's pointing at each of the positions 1 through 12. If you walk past his office when the door's open, you may be able to catch a quick glance at the clock. Suppose you do. Intuitively: you know that upon glancing at the clock, you should be able to tell roughly—but not exactly—where it's pointing.

Here's one way to sharpen this description into a prior frame. You (initially) should be sure that: if it's pointing at n, then upon glancing at it you should be sure that it's pointing somewhere between $n-1$ and $n+1$. This generates the prior frame in Figure 2. The set of worlds is $W = \{1, 2, ..., 12\}$. The accessibility relation E is encoded

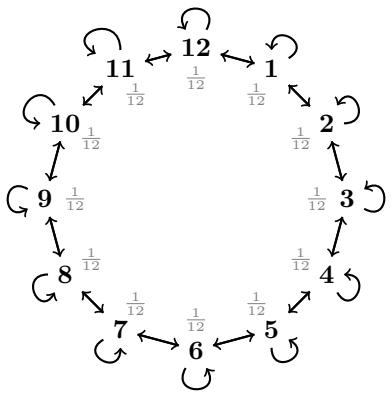

Figure 2: The Unmarked Clock

in the arrows, so for example $E_1 = \{12, 1, 2\}$, $E_2 = \{1, 2, 3\}$, etc. The prior π is encoded in the faded '$\frac{1}{12}$' written next to each world. Since $P_i(\cdot) = \pi(\cdot|E_i)$, it follows that at each world i you should be equally ($1/3$) confident that you're at $i-1$, i, and $i+1$—for example, $P_2(1) = P_2(2) = P_2(3) = 1/3$.

This model leads to higher-order uncertainty: since what you should think depends on where the hand is pointing, and you should be unsure where it's pointing, you should thereby be unsure what you should think. For example, at 2 you should be $1/3$ confident in each of worlds 1, 2, and 3; meanwhile, at 3 you should assign credence 0 to being at 1. Therefore at 2 you should be unsure whether your confidence that you're at world 1 should be $1/3$ or 0: $P_2(P(1) = 1/3) > 0$ and $P_2(P(1) = 0) > 0$.

This is a version of the case presented in Williamson (2014); see Salow (2018b); Ahmed and Salow (2018). It's an intuitive-enough description, and an elegant toy model. So far, it illustrates two of our three methods: describing a simple scenario, and drawing up a toy formal model of it. But it also illustrates the need for the third method—motivating our theory with general principles—for if the model represents a

genuine possibility, a variety of very surprising phenomena can be rational.

First, the model allows Improbable Certainties.[22] Let $\mathit{Early} = \{1, 2, 3\}$ be the proposition that it's pointing at 1, 2, or 3. Then at world 2 you should be certain it's early, but everywhere else you should be unsure of this: $[P(\mathit{Early}) = 1]$ is true at 2, and false everywhere else. Since at 2 you should be only $1/3$ confident that you're at 2, it follows that if it's pointing at 2 then you should be sure that it's early, yet be only $1/3$ confident that you should be sure of this: $[P(\mathit{Early}) = 1] \wedge [P(P(\mathit{Early}) = 1) \leq 1/3]$ is true at 2. So you should think to yourself, "It's early, but I probably shouldn't be sure of that." It follows that you should also think to yourself, "Supposing it's not rational for me to be sure it's early, it's (still) definitely early": $P_2(\mathit{Early} | P(\mathit{Early}) < 1) = 1$.[23]

Second, the model allows Confidence-Akrasia. Let $\mathit{Odd} = \{1, 3, 5, 7, 9, 11\}$ be the proposition that it's pointing at an odd number. At 2 you should be $2/3$ confident of *Odd*, since two of the three possibilities you should leave open are 1 and 3. But, for parallel reasons, at each of 1 and 3 you should be only $1/3$ confident in *Odd*. Thus at 2 you should be $2/3$ confident in the conjunction "the hand is pointing at an odd number, but I shouldn't be $(2/3)$ confident of that": $P_2(\mathit{Odd} \wedge [P(\mathit{Odd}) < 2/3]) \geq 2/3$.

Third, this model allows Effacing Evidence. Note that the reasoning from the previous paragraph applies at *every* world: at every *Odd*-world, $[P(\mathit{Odd}) = 1/3]$ is true, and at every $\neg \mathit{Odd}$-world, $[P(\mathit{Odd}) = 2/3]$ is true. Thus the biconditionals $\mathit{Odd} \leftrightarrow [P(\mathit{Odd}) < 1/2]$ and $\neg \mathit{Odd} \leftrightarrow [P(\mathit{Odd}) > 1/2]$ are true at all worlds. Since at each world you should be sure that you're at one of the worlds, it follows that at each world you should be sure of these biconditionals: $[P(\mathit{Odd} \leftrightarrow [P(\mathit{Odd}) < 1/2]) = 1]$ and $[P(\neg \mathit{Odd} \leftrightarrow [P(\mathit{Odd}) > 1/2]) = 1]$ are both true at all worlds. In other words, you should be certain that the rational credences are pointing in the wrong direction about *Odd*—certain, for instance, that credence $1/2$ in *Odd* is more accurate than the rational credence in *Odd*, whatever it is (Horowitz 2014; Ahmed and Salow 2018; Salow 2018b; Dorst 2020).

Because of this, this clock model leads to radical failures of the **Value of Evidence**: the idea that you should always expect (free) evidence to help you make better decisions (Salow, this volume). For example, if you had to decide whether to bet at fair odds on *Odd*, against *Odd*, or abstain from betting either way, then you should be certain (both before and after glancing at the clock) that making the rational decision after glancing at the clock—whatever it is—will lead you to take the wrong bet and so lose money. Thus you should prefer to simply take no bet, rather than to look at the clock and take the option rationalized by your evidence (Ahmed and Salow 2018; Dorst 2020).

At this point, you might start to think that something has gone wrong—we've ended up in a sort of paradox. Could it really be that in deciding whether to bet on whether

[22] This, in fact, is why Williamson (2014) introduced it—though his focus was on knowledge.

[23] This is a variant on what Dorst (2019a) calls "abominable conditionals"; it arises whenever there are **Positive Access** failures—cases in which $P(q) = 1$ but $P(P(q) = 1) < 1$.

the hand is pointing at an odd number, you should prefer *not* to look at the clock? If want to avoid this result, we need to formulate general principles about the connection between first- and higher-order uncertainty that explain what's gone wrong with our toy model.

3 The Answers

Such principles can be thought of as (partial) theories of higher-order uncertainty, delineating what constraints must be met by a model for it to describe a genuine rational scenario. To defend such a principle, we need to do at least three things:

(1) Explicitly formulate and motivate the principle.
(2) Show that it is *nontrivial*, in the sense that it allows higher-order uncertainty.
(3) Show that it is *potent*, in the sense that is rules out (some) paradoxical scenarios.

As mentioned above, step (2) often leads to surprises—our first group of principles all (non-obviously) rule out higher-order uncertainty:

Rational Reflection: $P(q|P(q) = t) = t$ (For all q, t)
If you learn that the rational credence in q is t, adopt credence t in q.[24]

Estimate Enkrasia: $P(q) = \mathbb{E}[P(q)]$ (For all q)
Your credence in q should equal your best estimate of the rational credence in q.[25]

Stay Away from the Edge:
If $P(P(q) \leq t) = 1$ and $P(P(q) < t) > 0$, then $P(q) < t$ (For all q, t)
You should not have a credence on the edge of the range of credences you think might be rational.[26]

No Intentionally Biased Inquiry:
Rational inquiries are not biased in favor of any given proposition.[27]

These principles are all very natural; why do they turn out to be so strong? Focus on Reflection. When the rational credence function has higher-order uncertainty, it's not

[24]Principles like this are discussed in Gaifman (1988); Samet (2000); Christensen (2007b, 2010b); Williamson (2008); Roush (2009, 2016); Elga (2013); Mahtani (2017); Gallow (2019a). See Samet (2000), Elga (2013), Williamson (2014), Gallow (2017), and Dorst (2020) for triviality results.

[25]$\mathbb{E}$ here is the mathematical expectation with respect to P, i.e. $\mathbb{E}[P(q)] := \sum_t P(P(q) = t) \cdot t$. Principles like this are discussed in Samet (2000); Christensen (2010b); Pettigrew and Titelbaum (2014); Sliwa and Horowitz (2015); Skipper et al. (2018); Salow (2018b, 2019); Gallow (2019b). See Samet (2000) and Dorst (2019b) for triviality results.

[26]Principles like this are discussed in Christensen (2010b); Titelbaum (2010); Salow (2018b); and Fraser (2020), and related ones are discussed in Schultheis (2018); Hawthorne and Isaacs (2020). See Dorst (2021) for the triviality result.

[27]See Salow (2018b) for this principle and the triviality result; see also Titelbaum (2010); Das (2020b). The formulation of this principle (and hence the results) are contested—see Gallow (2019b).

certain what the rational credence function is. Thus *learning* about the rational credence function provides new information—and that information can *change* the rational credence function (Hall 1994; Elga 2013; Mahtani 2017; Dorst 2020).

For example, consider our Good/Bad Argument model. There are two possibilities: the argument is good (g) or bad (b). If it's good, you should be 0.8 confident it's good; if it's bad, you should be 0.4 confident it's good. Thus conditional on the rational confidence being 0.8 that it's good, how confident should you be that it's good? In other words, what is $P(g|P(g) = 0.8)$? Well, you know that it's good iff the rational credence that it's good is 0.8. Therefore if you learn that the rational credence is 0.8, you should become certain that it's good: $P(g|P(g) = 0.8) = 1$. The rational credence was 0.8 only because it didn't *know* the rational credence was 0.8; so when that higher-order uncertainty is removed, the rational credence changes.

Given this, Elga (2013) proposed a natural weakening of Rational Reflection. The idea is that when you learn about the rational credence function, you shouldn't simply adopt its opinions—rather, you should adopt the opinions it *would* have *were it to learn what you've learned*. He first formulates Rational Reflection this way:

> **Elga's Reflection:**: $P(q|P = \pi) \;=\; \pi(q)$[28] (For all q, π)
> Upon learning what the rational credence function is, adopt its opinions.

He notes that this principle straightforwardly crashes if the rational credence function has higher-order uncertainty, since $P(P = \pi|P = \pi) = \pi(P = \pi)$ only if $\pi(P = \pi) = 1$. In response, he proposes the natural refinement:

> **New Reflection:** $P(q|P = \pi) \;=\; \pi(q|P = \pi)$ (For all q, π)
> Upon learning what the rational credence function is, adopt the opinions it would have were it given that information.

Letting $\widehat{P}$ be the credence function that would be rational, were you informed of what the rational credence function was ($\widehat{P}_w(\cdot) := P_w(\cdot|P = P_w)$.), New Reflection is equivalent to the following:

> **HiFi:** $P(q) \;=\; \mathbb{E}[\widehat{P}(q)]$ (For all q)
> The rational credence in q equals the rational expectation of the credence that would be rational, were you informed of what the rational credence function was.[29]

New Reflection is a natural, well-motivated constraint, and it does not lead to higher-order triviality (see Pettigrew and Titelbaum 2014; Dorst 2020, 2019b). But it does face serious critiques. Some have argued that it is too strong (Lasonen-Aarnio 2015), while others have objected that it's too weak (Pettigrew and Titelbaum 2014; Dorst 2020,

[28]Here 'π' is a rigid designator for a particular probability function. Elga's Reflection is equivalent to Rational Reflection (Samet 2000), though things get more subtle when we have principles of the form $P(q|Q(q) = t) = t$ and $P(q|Q = \pi) = \pi(q)$, where $Q \neq P$; see Gallow (2017).

[29]See Dorst (2019b) for the equivalence to New Reflection, as well as Stalnaker (2019).

2019b). In particular, note that New Reflection is valid on our above model of the Unmarked Clock[30]—thus New Reflection permits Improbable Certainties, Confidence-Akrasia, Effacing Evidence, and radical failures of the Value of Evidence.

This observation suggests that New Reflection does not fully capture its own motivations. It says that when you learn *exactly* what the rational credence function is, you should adopt the credence it would adopt were it to learn what you learned. But there is a more general statement of that idea: if you learn *anything*, your credences should always be constrained by what you know about how the rational credence function would respond to what you learned. That doesn't hold in our Unmarked Clock model. For example, suppose you're at world 4. Then conditional on it being *Early* (i.e. $\{1,2,3\}$), you know the rational credence function leaves open that the hand is pointing at 2—and yet you should *not* leave open that it's pointing at 2.[31]

Here's a natural way to generalize New Reflection to avoid this. Letting $\boldsymbol{P^r(q)} := P(q|r)$ be the rational credence in q conditional on r (whatever it is):

> **Reaction:** If $P^r(l \leq P^r(q) \leq h) = 1$, then $l \leq P^r(q) \leq h$ (for all q, r, l, h)
> If you should be sure that the rational reaction to r is to adopt a credence in q in a given range, react to r by adopting a credence in that range.[32]

Our model of the clock fails to validate Reaction (see footnote 31).

So are we done? Can we say that what's wrong with our model of the clock—and all the paradoxical results that ensue from it—is that it violates Reaction, which is the proper generalization of New Reflection?

No. To say that, we'd need to show that our diagnosis (Reaction) won't allow the same paradoxes to arise in other shapes. And we can't show that, because Reaction *does* allow them to re-arise. Example: two sycophants—Sybil and Phan—share evidence, but Sybil is confident that Phan is rational, while Phan is confident that Sybil is rational. Here's a simple probability frame modeling their situation; s is the possibility where Sybil is rational, p is the possibility where Phan is:

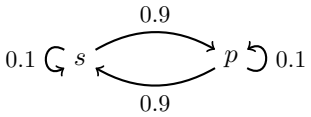

Figure 3: Sycophants

Reaction is valid on this frame, yet it allows Confidence Akrasia, Effacing Evidence, and radical failures of the Value of Evidence (Dorst 2020).

[30] Since the probability function P_w associated with each world w is unique, New Reflection is valid.

[31] $Early \subseteq [P(2|Early) \geq \frac{1}{3}]$, so $P_4(P(2|Early) \geq \frac{1}{3}|Early) = 1$, yet $P_4(2|Early) = P_4(2|\{3\}) = 0$.

[32] Reaction entails New Reflection (let $r = [P = \pi]$), but not vice versa. Although it may looks like Reaction follows from factivity (i.e. $[P(q) = 1] \rightarrow q$), it doesn't, for the antecedent of Reaction merely asserts that a *conditional* certainty (given r) holds: $P^r(l \leq P^r(q) \leq h) = 1$; not $P(l \leq P^r(q) \leq h) = 1$.

Upshot: if we want to rule out these puzzles, we need something that forces a correlation between what the evidence supports and truth—Reaction (or New Reflection) is not enough. Dorst (2020) argues that the following is a natural candidate:

> **Reliance:** $P^r(q|P^r(q) \geq t) \geq P^r(q)$ (For all q, r, t)
> Learning that it's rational to have confidence in q (given r) shouldn't lead you to lower your confidence in q (given r).

Reliance combined with Reaction is equivalent to the following:

> **Trust:** $P^r(q|P^r(q) \geq t) \geq t$ (For all r, q, t)
> Learning that it's rational to have confidence in q (given r) should lead you to become confident of q (given r).

Trust allows Living on the Edge and Unconfident Confidence, but rules out Improbable Certainties, Confidence Akrasia, and Effacing Evidence—and, within the class of prior frames, is equivalent to the Value of Evidence (Dorst 2020).

One challenge for Trust is that it is too strong. Although it allows significant amounts of higher-order uncertainty (Dorst 2020), it rules out certain models of "double bad cases" that are naturally motivated by epistemic asymmetries in externalist epistemology (Williamson 2019; Das 2020b).[33]

A different challenge for Trust is that it's too weak. Although it's equivalent to the Value of Evidence within prior frames, that connection doesn't hold in full generality: the Value of Evidence implies Trust, but it turns out the converse does not hold.[34] Does that mean that Trust needs to be strengthened?

Perhaps—but Ben Levinstein (2019) proves a result that partially alleviates this worry. The *unrestricted* Value of Evidence says that no matter what decision problem you face (betting, believing, etc.), you should expect the rational opinions to sanction decisions at least as good as your own. The ***Epistemic* Value of Evidence** says that no matter how you value accuracy, you should expect the rational credences to be at least as accurate as your own (Oddie 1997; Schoenfield 2016b). Levinstein shows, in effect, that Trust holds iff the epistemic value of evidence holds.

Nevertheless, we may still want a stronger principle. One worth exploring further is a "graded" form of introspection that weakens Access Internalism:

> **Graded Access:** If $P(q) \geq t$, then $P(P(q) \geq t) \geq t$ (For all q, t)
> If you should be confident of q, you should be confident that you should be.

[33] Using stochastic matrix notation (footnote 17), here's an example of a frame that has the structure Williamson and Das discuss: $\begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1 \end{pmatrix}$. See (Dorst 2020, Appendix A) for a reply to this case. Notably, Trust (and the Value of Evidence) do allow frames of this structure—they simply impose constraints on the probabilities. For example, they allow the following variant: $\begin{pmatrix} 1/5 & 1/5 & 1/5 & 2/5 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1 \end{pmatrix}$.

[34] Contra Dorst's (2020) Conjecture 7.3. Thanks to Branden Fitelson and Brooke Husic for help discovering this.

This is equivalent to the denial of Unconfident Confidence from §1, as well as to what Williamson (2019) calls "threshold-transfer" (compare Littlejohn 2019).[35] Although Williamson (2019) shows that this principle leads to higher-order triviality within prior frames, it seems to allow significant higher-order uncertainty in the context of probability frames—note that it is valid on our Good/Bad Argument frame.[36]

Summing up: there are several natural "joints" to be explored on the scale between weak theories that allow maximal amounts of higher-order uncertainty (and therefore minimal connections between higher- and first-order evidence) and those that allow minimal amounts of higher-order uncertainty (and therefore maximal such connections). Figure 4 (page 16) provides a map of these various positions and their known (⇒) or conjectured (?⇒) entailment relations. Bolded principles are relatively unexplored as standalone constraints, and (arguably) deserve further investigation.

4 The Applications

What can one *do* with these theories, once they are developed? Most obviously, they offer replies to our initial dilemma. For instance, Value and Trust will say that (i) is true—sometimes we can get misleading higher-order evidence—but that Confidence Akrasia is irrational, meaning that (ii) does not follow. More generally, the amount of allowed higher-order uncertainty will have bearings on a variety of further debates:

- Internalism vs. externalism, or the extent to which we have rational access to the requirements of rationality.[37]
- Whether and how you should defer to your (future) evidence.[38]
- Whether it's possible to rationally, intentionally bias your inquiries.[39]
- Whether evidence of evidence is evidence.[40]
- Whether we should always prefer to gather and use free evidence.[41]

[35] Threshold-transfer says that $[P(P(q) \geq t) \geq t] \rightarrow [P(q) \geq t]$. This is stronger than both Trust and the Value, which entail only a weakened version of the principle (Dorst 2020; Williamson 2019): $[P(P(q) \geq t) \geq s] \rightarrow [P(q) \geq t \cdot s]$.

[36] It is also valid on $\begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/4 & 1/2 & 1/4 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}$. What Williamson shows is that within prior frames, Graded Access comes very close to implying **Negative Access** (if $P(q) < 1$ then $P(P(q) < 1) = 1$). In prior frames this rules out higher-order uncertainty—but not so in probability frames.

[37] See Williamson (2000, 2008, 2014, 2019); Mahtani (2008); Smithies (2012, 2015); Cohen and Comesaña (2013); Lasonen-Aarnio (2013, 2014, 2015, 2019); Schoenfield (2016b); Das and Salow (2016); Salow (2018b, 2019); Das (2020a); Gallow (2019b); Dorst (2019b); Carr (2019a); Weatherson (2019).

[38] See Williamson (2008); Briggs (2009); Roush (2009, 2016); Christensen (2010b); Elga (2013); Pettigrew and Titelbaum (2014); Gallow (2017, 2019b); Schoenfield (2017a); Salow (2018b); Dorst (2020).

[39] See Titelbaum (2010); Salow (2018b); Dorst (2019c); Gallow (2019b); Das (2020b).

[40] See Eder and Brössel (2019); Williamson (2019); Salow (2018a).

[41] See Good (1967); Schoenfield (2016b); Ahmed and Salow (2018); Das (2020b); Dorst (2020); Levinstein (2019); Salow (this volume).

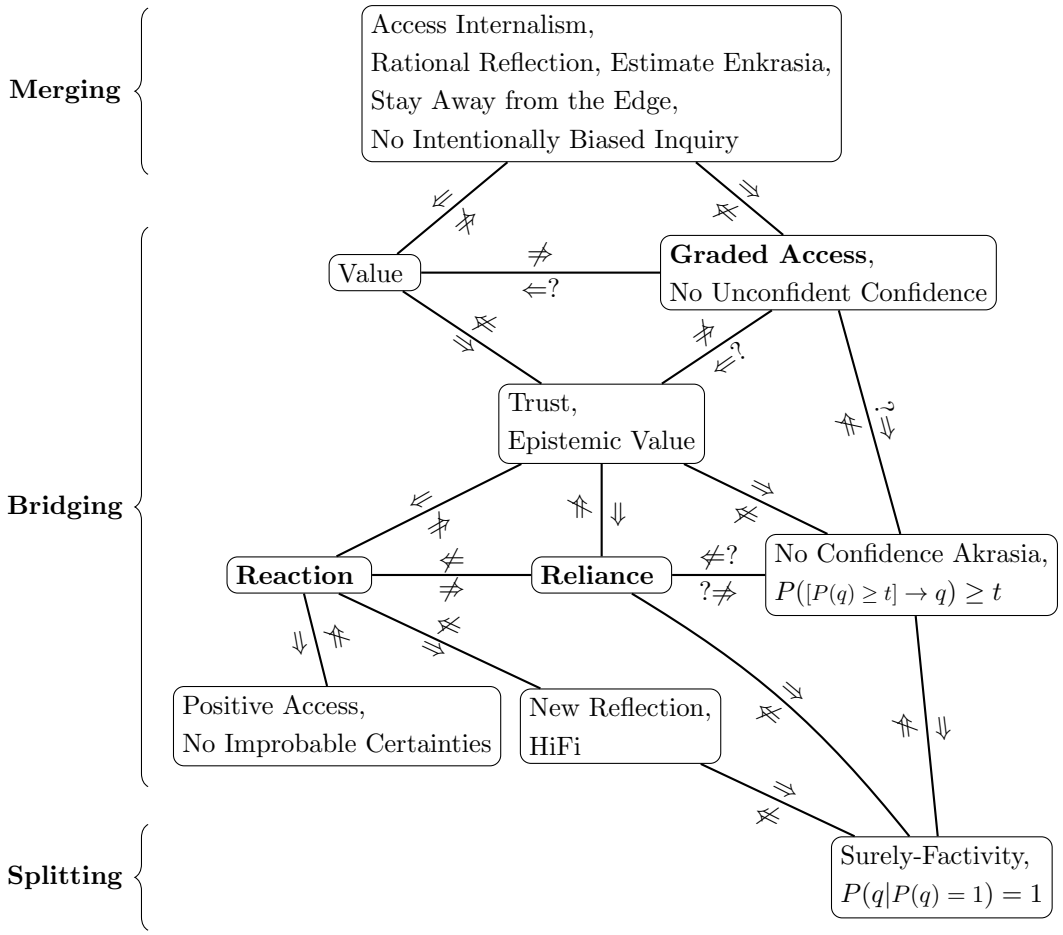

Figure 4: Map of Principles

- How to respond to peers who share our evidence but disagree with us.[42]

Here I'll focus on two further applications: ambiguous evidence and human (ir)rationality.

4.1 Ambiguous Evidence

Sometimes we get ambiguous evidence (Elga 2010)—for example: a man on the street removes two tubes of toothpaste and a live jellyfish from a bag; how confident should you be that the next item will be a tube of toothpaste? There's an ongoing debate about whether we should model the opinions of rational agents who face such evidence with precise probabilities, or *sets* of probability functions (Joyce 2010; White 2009a; Schoenfield 2012, 2017b; Konek 2019). But once we recognize the possibility of higher-order

[42]See Aumann (1976); Elga (2007); Christensen (2007c); Brössel and Eder (2014); Lederman (2015); Skipper et al. (2018); Dorst (2020).

uncertainty, there's another option: their evidence could rationalize precise credences along with higher-order uncertainty—there's some precise credence you should have that it'll be a tube of toothpaste, but you should be unsure what that rational credence is (Carr 2019b). Such higher-order uncertainty has strictly more structure than the set-theoretic representation, for in addition to the set of probability functions that you should leave open might be rational, it also includes various possible rational degrees of confidence over that set.[43]

Carr (2019b) makes a strong case for approach, and develops the foundations. The next step is to put it to work. One promising avenue is decision theory. It's notoriously difficult to find decision rules for sets-of-probabilities that behave well: many are subject to sure-loss arguments (Elga 2010; cf. Bradley and Steele 2014) or lead to violations of the Value of Evidence (Kadane et al. 2008; cf. Bradley and Steele 2016), for instance.

The higher-order-uncertainty approach may help, due to its richer structure. For example, two standard sets-of-probabilities rules are "conservative" and "permissive". The former says that an option is permissible iff it maximizes expected utility with respect to *all* probability functions in the set; the latter says: iff it maximizes expected utility with respect to *some* member of the set. In the higher-order uncertainty framework, the former is equivalent to saying that an option is permissible iff you should be *certain* that it maximizes expected utility; the latter is equivalent to saying: iff you should *leave open* that it maximizes expected utility.[44] But in this framework, there are many other options.

First, "maximize expected utility" remains an option, since you *should* have a precise credence function—you should just be unsure whether the one you have is rational. Thus the injunction "Choose an option o that maximizes $\sum_x P(U(o) = x) \cdot x$" makes sense—although often you will rationally be unsure which option maximizes this quantity, since you will be unsure what P is. On this way of doing things, provably there are theories of higher-order uncertainty that validate (or fail to validate) the value of evidence (see Geanakoplos 1989; Dorst 2020; Das 2020b). Alternatively, since higher-order probabilities induce expectations of expectations, another rule you could follow is "maximize *expected* expected utility", i.e. choose an option that maximizes the expectation of the expectation of utility.[45] Interestingly, although uncertainty about expected utility usually leads to uncertainty about expected expected utility, you will usually be *less* uncertain about the latter—so that if we keep iterating expectations you will

[43] Precisely, in addition to $\{\pi | P(P = \pi) > 0\}$, the rational credences P are also capture opinions about this set—e.g. $P(P = \pi_1) > P(P = \pi_2)$.

[44] Precisely, o_1 is permissible iff (conservative:) $P(\forall o_2 \colon \mathbb{E}[U(o_1)] \geq \mathbb{E}[U(o_2)]) = 1$; or (permissive:) $P(\forall o_2 \colon \mathbb{E}[U(o_1)] \geq \mathbb{E}[U(o_2)]) > 0$.

[45] For any random variable X, $\mathbb{E}[X] := \sum_x P(X = x)x$. Since in this framework $P(X = x)$ varies across worlds, it follows that $\mathbb{E}[X]$ is itself a random variable that varies across worlds, and therefore that $\mathbb{E}[\mathbb{E}[X]]$ is well-defined and potentially distinct from $\mathbb{E}[X]$ (Samet 2000; Williamson 2008). Thus "choose an option o that maximizes $\mathbb{E}[\mathbb{E}[U(o)]]$" makes sense.

eventually be certain of the ordinal rankings of options.[46] This implies, for instance, that decision rules of the form "iterate expectations until you are at least t-confident of some option that it maximizes (iterated) expected utility, and then do that" make sense and are worth exploring.

Notably, all of these approaches coincide with "maximize expected utility" when P is higher-order certain, but come apart when it is not.

4.2 Human (Ir)rationality

One of the motivations for studying higher-order uncertainty is that you and I often have it—we are often unsure whether we are thinking as we should. Thus as theories of rational higher-order uncertainty develop, one clear point of application is what they tell us about the (ir)rationality of human reasoners like us.

For example, Brian Hedden (2018) shows that given minimal assumptions about higher-order uncertainty, *hindsight bias*—the tendency to think that events were more predictable after you learn that they occurred—is often rational. In particular, this follows from two claims: (1) higher-order uncertainty is often rational, and (2) Reliance—the claim that what the evidence supports is correlated with truth—usually holds.

As another example, whenever higher-order uncertainty is rational it arguably follows that you can expect gathering more information to make it rational to raise your credence in some particular claim q.[47] For example, suppose you're about to be presented with an argument and you're currently 50-50 on whether it'll be a good or bad argument. Afterwards you'll be in the scenario describe by our Good/Bad Argument model in Figure 1. Then although your current rational credence is 50%, your current rational estimate for the future rational credence is $0.5(0.8) + 0.5(0.4) = 60\%$. In other words, you rationally expect that being presented with the argument should lead you to raise your credence that it's a good argument.

Interestingly, the Good/Bad Argument model validates the Value of Evidence—meaning it's the sort of evidence you should prefer to receive. As such, it raises the possibility of a rational form of *confirmation bias* (Nickerson 1998): the tendency to gather evidence that you expect to confirm a given opinion. This means that the rationality of higher-order uncertainty may help refine our understanding of empirical work on confirmation bias (Dorst 2019c).

[46] Precisely: a consequence of the convergence theorems for Markov chains (e.g. Bertsekas and Tsitsiklis 2008, §7.3) is that in most finite probability frames, for any X, there's a fixed point that the series $\mathbb{E}[X], \mathbb{E}[\mathbb{E}[X]], \mathbb{E}[\mathbb{E}[\mathbb{E}[X]]]$... will converge to at all worlds, meaning you rationally will become more and more confident of the values of these iterated expectations.

[47] See Titelbaum (2010); Salow (2018b); Das (2020b). This has been contested by Gallow (2019b).

5 Conclusion

Every now and then, philosophers come across a subtle phenomenon—usually dreamt up in a toy case—that turns out to be *everywhere* once you look for it. Arguably, higher-order evidence is such a phenomenon. Our lives are full of both self-doubt and its attendant effects: we are constantly wondering whether we are thinking as we should, and this exerts significant rational pressure on how we think. Thus the roots and importance of higher-order evidence go deep.

At the same time, the philosophical debate is young: many of the methods (§2), theoretical positions (§3), and applications (§4) are still being mapped. As such, higher-order evidence—and it's correlate, higher-order uncertainty—represents an excitingly open topic that criss-crosses the boundaries of formal, traditional, and applied epistemology. Much remains to be done.[48]

[48]Thanks to Maria Lasonen-Aarnio, Bernhard Salow, and Mattias Skipper for helpful comments on earlier versions of this chapter.

References

Ahmed, Arif and Salow, Bernhard, 2018. 'Don't Look Now'. *British Journal for the Philosophy of Science*, To appear.

Aumann, R, 1976. 'Agreeing to Disagree'. *The Annals of Statistics*, 4:1236–1239.

Bertsekas, Dmitri P and Tsitsiklis, John N, 2008. *Introduction to Probability*. Athena Scientific, second edition.

Bradley, Seamus and Steele, Katie, 2014. 'Should subjective probabilities be sharp?' *Episteme*, 11(3):277–289.

———, 2016. 'Can free evidence be bad? Value of information for the imprecise probabilist'. *Philosophy of Science*, 83(1):1–28.

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

Brössel, Peter and Eder, Anna-Maria A, 2014. 'How to resolve doxastic disagreement'. *Synthese*, 191(11):2359–2381.

Campbell-Moore, Catrin, 2016. *Self-Referential Probability*. Ph.D. thesis.

Carr, Jennifer Rose, 2019a. 'A modesty proposal'. *Synthese*, 1–21.

———, 2019b. 'Imprecise Evidence without Imprecise Credences'. *Philosophical Studies*, To appear.

Christensen, David, 2007a. 'Does Murphy's Law Apply in Epistemology? Self-Doubt and Rational Ideals'. In Tamar Szabó Gendler and John Hawthorne, eds., *Oxford Studies in Epistemology*, volume 2, 3–31. Oxford University Press.

———, 2007b. 'Epistemic Self-Respect'. *Proceedings of the Aristotelian Society*, CVII(3):319–337.

———, 2007c. 'Epistemology of Disagreement: The Good News'. *Philosophical Review*, 116(2):187–217.

———, 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*.

Cohen, Stewart and Comesaña, Juan, 2013. 'Williamson on Gettier Cases and Epistemic Logic'. *Inquiry*, 56(1):15–29.

Cresto, Eleonora, 2012. 'A Defense of Temperate Epistemic Transparency'. *Journal of Philosophical Logic*, 41(6):923–955.

Das, Nilanjan, 2020a. 'Externalism and Exploitability'. *Philosophy and Phenomenological Research*, To Appear.

———, 2020b. 'The Value of Biased Information'. *The British Journal for the Philosophy of Science*, To Appear.

Das, Nilanjan and Salow, Bernhard, 2016. 'Transparency and the KK Principle'. *Noûs*.

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

———, 2019b. 'Higher-Order Uncertainty'. In Mattias Skipper Rasmussen and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidence: New Essays*, 35–61. Oxford University Press.

———, 2019c. 'Why Rational People Polarize'. *Phenomenal World*.

———, 2020. 'Evidence: A Guide for the Uncertain'. *Philosophy and Phenomenological Research*, 100(3):586–632.

———, 2021. 'Be modest: you're living on the edge'. *Analysis*, To appear.

Eder, Anna-Maria A and Brössel, Peter, 2019. 'Evidence of Evidence as Higher-Order Evidence'. In Mattias Skipper and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidence: New Essays*, 62–83. Oxford University Press.

Elga, Adam, 2007. 'Reflection and Disagreement'. *Noûs*, 41(3):478–502.

———, 2010. 'Subjective probabilities should be sharp'. *Philosophers' Imprint*, 10(5):1–11.

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

Fraser, Rachel, 2020. 'Mushy Akrasia'.

Gaifman, Haim, 1988. 'A Theory of Higher Order Probabilities'. In Brian Skyrms and William L Harper, eds., *Causation, Chance, and Credence*, volume 1, 191–219. Kluwer.

Gallow, J. Dmitri, 2017. 'Local & Global Experts'.

———, 2019a. 'A Subjectivist's Guide to Deterministic Chance'. *Synthese*, To Appear:1–34.

———, 2019b. 'Updating for externalists'. *Noûs*, (November 2018):1–30.

Geanakoplos, John, 1989. 'Game Theory Without Partitions, and Applications to Speculation and Consensus'. *Research in Economics*, Cowles Fou(914).

Good, I J, 1967. 'On the Principle of Total Evidence'. *The British Journal for the Philosophy of Science*, 17(4):319–321.

Goodman, Jeremy, 2013. 'Inexact Knowledge without Improbable Knowing'. *Inquiry*, 56(1):30–53.

Greco, Daniel, 2014. 'A puzzle about epistemic akrasia'. *Philosophical Studies*, 161:201–219.

———, 2019. 'Justifications and Excuses in Epistemology'. *Nous*, To appear.

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

Hawthorne, John and Isaacs, Yoaav, 2020. 'Permissivism, Margin-for-Error, and Dominance'. *Philosophical Studies*, 1–18.

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

Hintikka, Jaako, 1962. *Knowledge and Belief*. Cornell University Press.

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

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

Hughes, George Edward, Cresswell, Max J, and Cresswell, Mary Meyerhoff, 1996. *A new introduction to modal logic*. Psychology Press.

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

Joyce, James M., 2010. 'A Defense of Imprecise Credences in Inference and Decision Making'. *Philosophical Perspectives*, 24(1):281–323.

Kadane, Joseph B, Schervish, Mark, and Seidenfeld, Teddy, 2008. 'Is ignorance bliss?' *The Journal of Philosophy*, 105(1):5–36.

Kappel, Klemens, 2019. 'Escaping the Akratic Trilemma'. In Mattias Skipper and Asbjørn Steglich-Petersen, eds., *Higher-order evidence: new essays*, 124–143. Oxford University Press.

Konek, Jason, 2019. 'Epistemic conservativity and imprecise credence'. *Philosophy and phenomenological research*.

Kripke, Saul A, 1963. 'Semantical analysis of modal logic i normal modal propositional calculi'. *Mathematical Logic Quarterly*, 9(56):67–96.

Lasonen-Aarnio, Maria, 2010. 'Unreasonable Knowledge'. *Philosophical Perspectives*, 24(1):1–21.

———, 2013. 'Disagreement and evidential attenuation'. *Nous*, 47(4):767–794.

———, 2014. 'Higher-order evidence and the limits of defeat'. *Philosophy and Phenomenological Research*, 8(2):314–345.

———, 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*, 144–171.

———, 2020. 'Enkrasia or evidentialism? Learning to love mismatch'. *Philosophical Studies*, 177(3):597–632.

Lederman, Harey, 2015. 'People with Common Priors Can Agree to Disagree'. *The Review of Symbolic Logic*, 8(1):11–45.

Levinstein, B. A., 2019. 'Accuracy, Deference, and Chance'.

Littlejohn, Clayton, 2018. 'Stop making sense? On a puzzle about rationality'. *Philosophy and Phenomenological Research*, 96(2):257–272.

———, 2019. 'Should we be dogmatically conciliatory?' *Philosophical Studies*, 177(5):1381–1398.

Littlejohn, Clayton and Dutant, Julien, 2019. 'Defeaters as Indicators of Ignorance'. In Mona Simion and Jessica Brown, eds., *Reasons, Justifications, and Defeat*. Oxford University Press.

Mahtani, Anna, 2008. 'Williamson on inexact knowledge'. *Philosophical studies*, 139(2):171–180.

———, 2017. 'Deference, respect and intensionality'. *Philosophical Studies*, 174(1):163–183.

Neta, Ram, 2019. 'The Puzzles of Easy Knowledge and of Higher-Order Evidence: A Unified Solution'. In Mattias Skipper and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidence: New Essays*, 173–188. Oxford University Press.

Nickerson, Raymond S., 1998. 'Confirmation bias: A ubiquitous phenomenon in many guises.' *Review of General Psychology*, 2(2):175–220.

Oddie, G, 1997. 'Conditionalization, cogency, and cognitive value'. *The British Journal for the Philosophy of Science*, 48(4):533.

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

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, 2018a. 'Evidence of evidence'.

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

———, 2019. 'Elusive Externalism'. *Mind*, 128(510):397–427.

Samet, Dov, 1999. 'Bayesianism without learning'. *Research in Economics*, 53:227–242.

———, 2000. 'Quantified Beliefs and Believed Quantities'. *Journal of Economic Theory*, 95(2):169–185.

Schoenfield, Miriam, 2012. 'Chilling out on epistemic rationality'. *Philosophical Studies*, 158(2):197–219.

———, 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.

———, 2017a. 'Conditionalization Does Not (In General) Maximize Expected Accuracy'. *Mind*, 126(504):1155–1187.

———, 2017b. 'The accuracy and rationality of imprecise credences'. *Noûs*, 51(4):667–685.

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

Sider, Theodore, 2010. 'Logic for philosophy'.

Skipper, Mattias, 2019a. 'Higher-Order Defeat and the Impossibility of Self-Misleading'. *Higher-Order Evidence: New Essays*, 189.

———, 2019b. 'Higher-Order Defeat and the Impossibility of Self-Misleading Evidence'. In Mattias Skipper and Asbjørn Steglich-Petersen, eds., *Higher-Order Evidence: New Essays, volume To appear*. Oxford University Press.

———, 2020. 'Does Rationality Demand Higher-Order Certainty?' *Synthese*, To Appear.

Skipper, Mattias, Steglich-Petersen, Asbjørn, and Bjerring, Jens Christian, 2018. 'A higher-order approach to disagreement'. *Episteme*, 15(1):80–100.

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

Smithies, Declan, 2012. 'Moore's paradox and the accessibility of justification'. *Philosophy and Phenomenological Research*, 85(2):273–300.

———, 2015. 'Ideal Rationality and Logical Omniscience'. *Synthese*, 192(9):2769–2793.

———, 2019. *The epistemic role of consciousness*. Philosophy of Mind.

Stalnaker, Robert, 2019. 'Rational Reflection, and the Notorious Unmarked Clock'. In *Knowledge and Conditionals: Essays on the Structure of Inquiry*, 99–112. Oxford University Press.

Tal, Eyal, 2018. 'Self-Intimation, Infallibility, and Higher-Order Evidence'. *Erkenntnis*, To appear.

———, 2020. 'Is Higher-Order Evidence Evidence?' *Philosophical Studies*, To Appear.

Titelbaum, M, 2019. 'Return to reason'. *Higherorder evidence. New essays*.

Titelbaum, Michael, 2015. 'Rationality's Fixed Point (or: In Defense of Right Reason)'. In Tamar Szabó Gendler and John Hawthorne, eds., *Oxford Studies in Epistemology*, volume 5, 253–292. Oxford University Press.

Titelbaum, Michael G., 2010. 'Tell me you love me: Bootstrapping, externalism, and no-lose epistemology'. *Philosophical Studies*, 149(1):119–134.

van Ditmarsch, Hans, Halpern, Joseph Y, van der Hoek, Wiebe, and Kooi, Barteld, 2015. *Handbook of Epistemic Logic*. College Publications.

Weatherson, Brian, 2019. *Normative externalism*. Oxford University Press.

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.

———, 2008. 'Why Epistemology Cannot be Operationalized'. In Quentin Smith, ed., *Epistemology: New Essays*, 277–300. Oxford University Press.

———, 2013. 'Response to Cohen, Comesaña, Goodman, Nagel, and Weatherson on Gettier Cases in Epistemic Logic'. *Inquiry*, 56(1):77–96.

———, 2014. 'Very Improbable Knowing'. *Erkenntnis*, 79(5):971–999.

———, 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.

Worsnip, Alex, 2018. 'The conflict of evidence and coherence'. *Philosophy and Phenomenological Research*, 96(1):3–44.

———, 2019. 'Can Your Total Evidence Mislead About Itself?' *Higher-Order Evidence: New Essays*, 298.