Fragmentation and logical omniscience

Adam Elga and Agustín Rayo*

November 17, 2020

(Main text length: 17 pages)

Abstract

It would be good to have a Bayesian decision theory that assesses our decisions and thinking according to everyday standards of rationality—standards that do not require logical omniscience (Garber 1983, Hacking 1967). To that end we develop a "fragmented" decision theory in which a single state of mind is represented by a family of credence functions, each associated with a distinct choice condition (Lewis 1982, Stalnaker 1984). On the resulting theory, rationality requires ordinary agents to be logically competent and to engage in trains of thought that increase the unification of their states of mind. But rationality does not require ordinary agents to be logically omniscient.

*Thanks to Diego Arana Segura, Alejandro Pérez Carballo, Ross Cameron, David Chalmers, Jonathan Cohen, Keith DeRose, Sinan Dogramaci, Cian Dorr, Kenny Easwaran, Hartry Field, Branden Fitelson, Peter Fritz, Daniel Hoek, Frank Jackson, Shivaram Lingamneni, Christopher Meacham, Patrick Miller, Molly O'Rourke-Friel, Michael Rescorla, Ted Sider, Mattias Skipper, Robert Stalnaker, Jason Stanley, Bruno Whittle, Robbie Williams; participants in the Corridor reading group (on three occasions), a graduate seminar session at Rutgers University, a Fall 2011 joint MIT/Princeton graduate seminar, and a Spring 2016 MIT/Princeton/Rutgers graduate seminar taught jointly with Andy Egan; audiences at the 2008 Arizona Ontology Conference, Brown University, Catholic University of Peru, CUNY, National Autonomous University of Mexico, Ohio State University, Syracuse, University, University of Bologna, UC Berkeley, UC Riverside, UC Santa Cruz, University of Connecticut at Storrs, University of Graz, University of Leeds, University of Paris (IHPST), University of Oslo (on two occasions), University of Texas at Austin, Yale University, MIT, and Rutgers University. The initial direction of this paper was enormously influenced by conversations with Andy Egan. Elga gratefully acknowledges support from a 2014-15 Deutsche Bank Membership at the Princeton Institute for Advanced Study.

1 *Standard decision theory is incomplete*

> Professor Moriarty has given John Watson a difficult logic problem and has arranged for a bomb to explode unless Watson gives the correct answer by noon. Watson has never thought about that problem before, and even experienced logicians take hours to solve it. It is seconds before noon.
>
> Watson is then informed that Moriarity has accidentally left the answer to the problem on a note attached to the back of the bomb. Watson's options are to look at the note or to give an answer of his choice without looking at the note. Is it rationally permissible for Watson to look at the note?

The answer is elementary: it is rationally permissible.

Someone might object that only a logically omniscient agent could be fully rational, and therefore that Watson is required to be certain of the correct answer to the logic puzzle (and to give that answer). Even so, we hope the objector would agree that there is *a* sense in which, given Watson's limited cognitive abilities, it is rational, reasonable, or smart for Watson look at the note.[1,2]

Unfortunately, standard Bayesian decision theory (as it is usually applied) fails to deliver any sense in which it is rationally permissible for Watson to look at the note. For it represents the degrees of belief of an agent as a probability function satisfying the standard probability axioms. And on the usual way of applying these axioms to a case like Watson's, they entail that Watson assigns probability 1 to every logical truth, including the solution to Moriarity's logic problem.[3] But if Watson is certain of the solution from

[1]The Watson case is structurally similar to the "bet my house" case from Christensen (2007, 8–9). For arguments that seek to differentiate between "ordinary standards of rationality" (according to which logical omniscience is not required) and "ideal standards" (according to which it is), see Smithies (2015).

[2]Compare: even an objective Bayesian who counts some prior probability functions as irrational might have use for a decision theory that says what decisions are rational, given a particular (perhaps irrational) prior.

[3]For important early discussions of how the assumption that logical truths gets probability 1 makes trouble for decision theory, see Savage (1967, 308) and Hacking (1967). For

the start, the small effort required to look at the note would offer no gain. It would therefore follow from the rule of maximizing utility that it is not rationally permissible for him to look at the note.

The above example shows that standard Bayesian decision theory is incomplete. Can we develop a decision theory that correctly assesses the behavior of agents like Watson, who satisfy everyday standards of rationality without being logically omniscient?[4] And can we model the logical and mathematical thinking of such agents?

2 *Rationality requires at least partial coherence*

Getting a decision theory for agents who fail to be logically omniscient might seem simple: just remove the assumption of logical omniscience from decision theory! To do that, one might (1) take credence functions to be defined over an appropriate set of fine-grained entities (which we will for

more recent discussions, see for example Easwaran (2011, §2.2), Seidenfeld et al. (2012), and Dogramaci (2018b).

There are many varieties of Bayesianism and many decision theories, and hence many theories that are reasonable candidates for the name "standard Bayesian decision theory". Bayesian frameworks such as Earman (1992) and Kaplan (1996) take the objects of probability to be sentences, and adopt axioms that entail that logical truths get probability 1. Other frameworks take the objects of probability to be the members of an algebra of propositions (which are often taken to be sets of elementary possibilities) and assign probability 1 to necessary propositions (Jeffrey 1965, Joyce 1999, Lewis 1981, Savage 1954). On the standard way of applying such frameworks to the case of Watson, what determines what Watson should do is his opinion on the solution to the puzzle. On the sentential approach, sentences that express the solution to the puzzle are logical truths, and hence are always assigned probability 1. On the propositional approach, any proposition expressing the solution to the puzzle is a necessary truth, and hence is always assigned probability 1.

[4]We take this challenge to be an interesting part (though not the only part) of what is sometimes called the "problem of logical omniscience" (Hintikka 1975). Much of the existing work on the problem of logical omniscience does not address the topic of the present paper: the problem of giving a probabilistic decision theory for logically non-omniscient agents. Instead, much work focuses on adapting Kripke models for all-or-nothing belief or knowledge to accommodate omniscience failures. On the connection between the problem of logical omniscience in the all-or-nothing belief and probabilistic belief contexts, see Cozic (2006, 56). For a survey of approaches to the problem of logical omniscience, see Halpern and Pucella (2011). Works that address probabilistic versions of the problem of logical omniscience include Dogramaci (2018b), Gaifman (2004), Garber (1983), Hacking (1967), Hoek (2019), Lipman (1999), Parikh (2008), Pettigrew (2020), Seidenfeld et al. (2012), Skipper and Bjerring (2020).

the moment assume to be sentences), but (2) refrain from imposing the usual coherence requirements (such as the requirement that logically equivalent sentences get the same credence).[5]

There are a few theoretical questions that need to be answered in order to get this approach off the ground (see Appendix A). But suppose for the sake of argument that they have been addressed and that we have a decision theory according to which imperfect agents like Watson have incoherent credence functions. In such a framework, Watson might have low credence in a sentence that expresses the solution to Moriarity's logic puzzle, even if that sentence is a logical truth. If so, we get the desired result that it is rational for Watson to look at the note.

Still there would be something missing. For although everyday standards of rationality allow for some failures of logical omniscience, not just anything goes. For example, assuming Watson understands the logical connectives, it would be irrational for him to assign more credence to "it is sunny and windy" than to "it's sunny". And the same would go for assignments that violate other obvious logical entailments. But were we to discard the standard probabilistic coherence assumptions altogether, nothing would rule out such assignments.[6]

A bit of terminology: For q a (perhaps incoherent) credence function that assigns a real number to each sentence of an appropriate language, say that q *respects* an entailment from A to B when $q(A) \leq q(B)$.[7] The above discussion suggests that some entailments—call them the "obvious" ones—are such

[5]Prior work in this spirit includes Caie (2013), Dogramaci (2018a,b), Gaifman (2004), Hacking (1967), Staffel (2015), Zynda (1996). Related work that allows for incoherent previsions (betting rates) defined over a space of elementary possibilities includes Schervish et al. (2003), Seidenfeld et al. (2012).

[6]Works recognizing the need for coherence requirements on agents that fall short of full coherence include Hacking (1967), Gaifman (2004), Bjerring (2013), Jago (2014), Elga and Rayo (2017), Dogramaci (2018b), and Skipper and Bjerring (2020). The papers by Bjerring, Jago, Skipper and Bjerring, and our own earlier work also include variants of the argument below, which shows that it's hard to respect obvious entailments without respecting all entailments.

[7]We use "q" rather than "P" to emphasize that q is defined over sentences rather than propositions, and that q is not assumed to satisfy the ordinary coherence axioms. Later in the paper we will work with probability functions defined over propositions understood as sets of possible worlds and will use "P" to refer to such functions.

that all rational agents have credence functions that respect them. So what seems to be needed is a way of requiring a credence function to respect all obvious entailments without requiring it to respect any non-obvious ones.

Unfortunately, there is no such way.

To see why, let $\langle A_1, A_2, \ldots, A_n \rangle$ be any chain of entailments (sequence of sentences, each of which entails the next). Then if a credence function q respects each successive entailment in the chain, it also respects the one-step entailment from A_1 to A_n. (Proof: by the transitivity of the less-than relation amongst real numbers, $q(A_1) \leq q(A_2) \leq \cdots \leq q(A_n)$ entails $q(A_1) \leq q(A_n)$.)

This observation leads to trouble because, on any interesting way of spelling out "obvious", chaining together obvious entailments can result in a non-obvious one. For example, many proofs and mathematical calculations lead to non-obvious conclusions by way of obvious steps. Moral: there is no way to require a credence function to simultaneously respect all obvious entailments without also requiring it to respect some non-obvious entailments.

So a challenge remains: how can we require that an agent's credences reflect logical competence without requiring the agent to be logically omniscient?

In the next few sections we develop an answer to this challenge. The crucial move will be to drop the assumption that a subject's decision-theoretic state is represented by a single credence function. We will instead adopt *fragmented decision theory*, according to which an agent at a time may have multiple credence functions.[8] Let us explain.

[8]We do not claim that fragmented decision theory is the only attractive way out of the dilemma—just that it is one approach worth investigating. A salient alternative approach (fruitfully developed in Skipper and Bjerring (2020, §3)): don't posit fragmentation, but instead impose novel constraints on how an agent's credence function changes over time. Articulating the relevant constraints requires a substantial bit of theorizing since (as we have seen) simply discarding the coherence conditions does not produce a theory that imposes strong enough logical competence constraints. But that is not intended as a criticism. Indeed, our own proposal could be adapted to develop adequate constraints for a dynamic account.

3 A decision-theoretic state can be represented by multiple credence functions

Fragmented decision theory is based on the idea—pioneered by Lewis (1982, 436) and Stalnaker (1984, Chapter 5)—that the across-the-board notion of information possession should be replaced with a notion of information accessible *relative to a condition*.

To see how this idea might be motivated, note that information can be represented in a way that makes it accessible for some purposes, but inaccessible for others (Stalnaker 1991, 437–438). Consider, for example, a pair of crossword-puzzle solvers trying to fill in the blanks below to complete a word of English:

__ __ __ __ M T

The first puzzlist fills in just the right letters. The second scratches his head and leaves the puzzle blank. Suppose further that each puzzlist knows that *dreamt* is a word of English, and knows how to spell it. Indeed, each puzzlist realizes from the start that *dreamt* is a word of English that ends in MT.[9]

So why is one puzzlist disposed to fill in the blanks with DREA, while the other is disposed to gnash his teeth, curse, and fill in nothing?

We suggest that both puzzlists possess the information they need to fill in the blanks, but that the conditions relative to which they have access to this information are different. Let D be the set of worlds in which *dreamt* is a word of English spelled D-R-E-A-M-T. Both puzzlists have access to D for the purpose of using "dreamt" in a written essay. And they both have access to D for the purpose of answering the question "Is 'dreamt' a word of English ending in MT?". But for the purpose of filling in the blanks in "__ __ __ __ M T", only the first puzzlist has access to D.

So if we'd like to represent the difference between the two puzzlists, our representation of each of them need not specify what information he possesses, period. It can instead say what information he or she has access

[9]Similar examples include the "deny" example in Powers (1978, 341), the factoring example from Stalnaker (1991, 438), the "Do geese see god?" example from Crimmins (1992, 246), and the "Iceman" example from Egan (2008, 51).

to for what purposes.[10,11]

For example, we might represent the struggling puzzlist's decision-theoretic state (in the context of her efforts to solve the crossword puzzle) with the following sort of table, which we shall call an *access table*:[12]

[10]In an illuminating paper on the role of the organization of memory in human reasoning, Cherniak (1983, 166) notes that creatures with mental organizations anything like ours constantly face a hard computational problem: quickly accessing memories relevant to their current situation. Cherniak convincingly argues that there is therefore a practical need for a small short-term memory store that supports fast—but not exhaustive—searching and consistency checking. It is to be expected that the heuristics underlying such searches will operate differently in different circumstances, and hence that different information will be accessible in different circumstances. Indeed, the necessity for heuristics that narrow memory searches was already recognized by Hume: "as the production of all the ideas to which [a] name may be applied, is in most cases impossible, we abridge that work by a more partial consideration, and find but few inconveniences to arise in our reasoning from that abridgement" (Hume 1738, 21, as cited in Cherniak 1983, 176).

[11]In representing the decision-theoretic state of an agent using access tables, we do not mean to propose that the structure of an access table maps in any direct way to an agent's mental representations. Instead we remain neutral on the psychological realization of the states we model, just as a proponent of classical Bayesianism might remain neutral on the psychological realization of credence functions. For further discussion, see Elga and Rayo Forthcoming.

[12]The suggestion in Stalnaker (1984) that logical omniscience failures can be understood in terms of fragmented belief states was the core motivation for the present model. Braddon-Mitchell and Jackson (2007, 199–200) also uses fragmented coarse-grained belief states to accommodate failures of logical omniscience. Yalcin (2008, Ch. 3), Yalcin (2015), Yalcin (2016), and Hoek (2019) develop that same suggestion, proposing elegant models on which all-or-nothing belief is relative to questions, understood as partitions of logical space. The treatment of logical omniscience failures in those works uses privileged partitions to represent which propositions are accessible to an agent, and so differs from the present treatment. (See especially Yalcin (2016, n. 26).) Egan (2008) endorses a treatment of fragmented credences similar to the present one and interestingly suggests that mental fragmentation might be practically indispensable for agents with perceptual belief forming mechanisms anything like human ones—mechanisms that are less than perfectly reliable but which nevertheless produce immediate belief in certain circumstances. For more recent discussion of fragmented mental states, see the essays collected in Borgoni et al. (forthcoming).

Choice condition	*Accessible information*
working on puzzle, *dreamt* salient	P_1
working on puzzle, *dreamt* not salient	P_2
[more conditions]	[information accessible relative to those conditions]

Each row of an access table consists of a type of choice condition together with an associated probability function.[13] Each probability function, denoted with a subscripted "P", is assumed to be coherent and defined on sets of possible worlds. (In contrast, the functions denoted by "q" in section 2 were not assumed to be coherent and were defined on sentences.)

In the struggling puzzlist's access table, P_1 is a probability function that assigns high credence to the set of worlds in which *dreamt* is a word of English spelled D-R-E-A-M-T, and P_2 is a credence function that assigns low credence to that set. Accordingly, the table reflects that the puzzlist's dispositions factor into two natural components: one component associated with situations in which the word "dreamt" has been made salient, and another in which it has not.[14]

[13]What settles which choice conditions figure in a subject's access table? A satisfactory answer would in normal cases deliver a set of choice conditions that strikes a good balance between two competing requirements: (1) the choice conditions are fine-grained enough to capture as many of the subject's rational dispositions as possible (on an everyday understanding of rationality that does not require logical omniscience); and (2) the choice conditions are coarse-grained enough to ensure that the resulting access table delivers a reasonably systematic description of the subject's rational dispositions. For the worry that if choice conditions are individuated to finely, an access table might become a mere laundry-list of overly specific dispositions (and so fail to provide useful explanations of behavior), see Norby (2014, §2), Hoek (2019, 137–139), Quilty-Dunn and Mandelbaum (2018, 2358–2359). For further discussion, see Elga and Rayo (Forthcoming).

[14]The above treatment has been simplified in several respects:

First, a more realistic access table would include choice conditions specific enough to represent other respects in which Watson's decision theoretic state is fragmented.

Second, in a fuller treatment we would simultaneously consider fragmented credences and fragmented values. In particular, each row of an access table might consist of a type of choice condition and a pair $\langle P, V \rangle$ of a probability function and a value function. This would allow for fragmented values in addition to fragmented credences, and also would

Can we assess the actions of a fragmented subject for rationality? Stalnaker (1984, 85) suggests that we can: "All of my actions may be rational in that they are directed toward desired ends and guided by coherent conceptions of the way things are even if there is no single conception of the way things are that guides them all." Stalnaker's idea is that even if an agent's belief state is incoherent, it can be thought of as a composite of individually coherent components. On this way of thinking, we can assess an action for rationality relative to a particular component.

For example, the puzzlist's decision-theoretic state does not deliver a single coherent conception of the way things are. But each row of her access table does correspond to such a conception. So we can characterize actions that are rational for her (and for other fragmented agents) as follows:

FRAGMENTED CHOICE RULE A subject in choice condition c should act so as to maximize expected utility relative to P_c, where P_c is the probability function associated with c in the subject's access table.[15]

This rule generalizes the standard rule of utility maximization since it coincides with that rule in the special case of a perfectly coherent agent—an agent whose access table has only one row.

4 Logical competence is a global feature of access tables

With this as our background, let us return to the challenge raised in section 2. The challenge was to find a way of requiring an agent to respect all obvious

allow for hybrid fragmented states that mix together the two types of fragmentation.

Third, in a more sophisticated version of the proposal each entry in the second column of an access table would be a *set* of $\langle P, V \rangle$ pairs. That would allow for agents who have imprecise or indeterminate probabilities and values.

Fourth, we do not claim that the above table would be suitable for representing the puzzlist's decision-theoretic state in arbitrary contexts. For all we have argued here, different theoretical contexts might call for different access tables. We remain neutral on the question of whether these different tables could be consolidated into a single "super table".

[15]The exact manner in which a probability function and a value function determine expected utilities for options depends on the flavor of (standard) decision theory that one uses as a base. Leading options include causal decision theory (Gibbard and Harper 1978, Joyce 1999, Lewis 1981, Savage 1954, Stalnaker 1981) and evidential decision theory (Ahmed 2014, Jeffrey 1965). Since disputes between such theories and their competitors are orthogonal to present concerns, we remain neutral here.

logical entailments without requiring her to respect non-obvious ones. Now, if an agent's decision-theoretic state is represented by a single credence function, then the natural way to require the agent to respect an entailment is to require her (unique) credence function to respect it. As we saw in §2, this approach runs into trouble because forcing a single credence function to simultaneously respect all obvious entailments also forces it to respect many non-obvious ones.

But if an agent's decision-theoretic state is represented by multiple credence functions—if it is represented by an access table—then another approach becomes available. One can require the agent to respect an entailment by requiring her access table to respect that entailment *at an appropriate row*.

Recall that in an access table, each row contains a (fully coherent) probability function. Such probability functions must respect set-theoretic connections between propositions, understood as sets of possible worlds. But they need not assign probabilities to the truth of sentences in a way that respects logical entailments among those sentences. That is because there are possible worlds at which the connectives fail to satisfy the standard truth tables.[16]

Consider, for example, the entailment from "$S\&W$" to "S". (Here and below we let "S" mean that it is sunny and "W" mean that it is windy.) Logically competent agents should respect this entailment. To do so, it is enough that they possess some logical information: that any world in which "$S\&W$" is true is also a world at which "S" is true. But logically competent agents need not be required to have access to such information relative to all conditions—only some. Which ones? A natural answer: conditions in which "S" and "W" are salient.

These considerations suggest a general requirement for logically competent agents: relative to a choice condition, the agent's information should include all obvious entailments among sentences that are built directly from sentences salient at that condition.

On the resulting picture, one cannot be logically competent unless every

[16]For appeals to semantically deviant worlds in modeling mathematical belief, see Stalnaker (1984, Ch. 5), Stalnaker (1986). For critical discussions of such appeals, see Field (1986b), Field (1978, 34–35), Field (1986a), Soames (2009), Speaks (2006, 448–450), Williamson (2016, n. 1).

obvious entailment is reflected in one's access table. But logical competence does not thereby collapse into logical omniscience. That is because a competent agent's logical information will typically be distributed throughout her access table, with different batches of information accessible in different circumstances. Logical competence is thus a global feature of the agent's entire access table rather than a local feature of any particular credence function.

A similar point can be made about mathematical competence. Consider Gottlob Frege, shortly before he read Russell's fateful letter informing him of the inconsistency of his mathematical system (Russell 1902). On the one hand, we want to be able to say that Frege understood his own mathematical vocabulary, along with the background logic. On the other hand, we want to do justice to the fact that he did not realize his system entailed a contradiction.

One natural approach is to model Frege as respecting every obvious entailment that is warranted by his axioms and rules of inference, while failing to respect certain non-obvious entailments (including the entailment Russell discovered). As we have seen, there is no way of satisfying both of these conditions at once if we represent his decision-theoretic state using a single credence function. But if we represent his decision-theoretic state as an access table, we can model his understanding of the relevant mathematical vocabulary as a global feature of that table, with different batches of semantic information distributed across different rows. For example, we can model Frege so that the crucial information highlighted by Russell's letter—that Frege's Basic Law V entails Russell's (blatantly inconsistent) instance—is only accessible relative to choice conditions in which that instance is salient.

(But isn't this an overly linguistic approach to logical and mathematical ignorance? For example, didn't Frege have mistaken beliefs about the nature of *extensions*, not just mistaken beliefs about *language*? Furthermore, how does an agent's access table connect to what the agent *believes* and *understands*? To avoid digressing, we address these questions in Appendix B.)

5 Local Booleanism models Boolean competence

This section gives a simple realization of the picture of logical and mathematical competence outlined in the preceding section. We make two sim-

plifications. First, we restrict our attention to the special case of Boolean entailments: entailments guaranteed by the meanings of the Boolean connectives. Second, we let an entailment count as "obvious" in a given choice condition if it is guaranteed by the meanings of the Boolean connectives as they apply to sentences that are salient in that condition. For example, take a condition in which "S" and "W" are both salient. Relative to such a condition, the entailment from "S&W" to "S" counts as obvious because it is guaranteed by the meaning of "&", as it applies to "S" and "W".

Given these simplifications, it is natural to impose the following requirement on agents who understand the Boolean connectives: relative to each choice condition, the available information should include that the Boolean connectives behave standardly with respect to sentences salient at that condition.

To state the requirement precisely, assume that each agent has an *obviousness function*: a function $\mathcal{O}$ that assigns to each choice condition c the logical information $\mathcal{O}_c$ that is available as obvious to the agent in c.[17] We capture the fact that $\mathcal{O}_c$ is available as obvious at c by assuming that $P_c(\mathcal{O}_c) = 1$, where P_c is the probability function corresponding to c in the agent's access table. We then say that $\mathcal{O}$ is *locally Boolean* just in case, for each choice condition c, $\mathcal{O}_c$ entails that the logical connectives behave standardly with respect to every sentence that is salient in c.[18]

This allows us to express the following constraint on logical competence:

CONNECTIVES When an agent understands the Boolean connectives, her

[17] In other words, for each choice condition c: $\mathcal{O}_c$ is the set of possible worlds compatible with whatever logical information is available as obvious to the agent relative to c.

[18] For the Boolean connectives to behave standardly with respect some sentences is for the following conditions hold whenever A and B are among those sentences:

B0 $\top$ is true,

B1 $\ulcorner\neg A\urcorner$ is true iff A is not true,

B2 $\ulcorner A \vee B\urcorner$ is true iff either A or B is true,

B3 $\ulcorner A\&B\urcorner$ is true iff both A and B are true.

$\top$ is a zero-place connective that is true as a matter of logic. It is included here for technical convenience and could be replaced by a tautology whose truth is available as obvious to the subject with respect to any choice condition.

obviousness function should be **locally Boolean**.

In this (admittedly simple) model, we have arrived at a precisely specified requirement of logical competence that falls short of logical omniscience.

6 Local Booleanism is equivalent to local sentential coherence

(This semi-technical section may be skipped without loss of continuity.)

Imposing CONNECTIVES falls short of requiring full logical omniscience. But what exactly does it require? And how does that requirement relate to more familiar axioms for probabilistic coherence?

It is useful to start by considering "global" versions of these questions. Say that an obviousness function $\mathcal{O}$ is *globally Boolean* if for each choice condition c, $\mathcal{O}_c$ contains only worlds at which the Boolean connectives behave standardly with respect to *all* sentences.

How severe a constraint is the requirement that an agent's obviousness function is globally Boolean? The answer is: extremely severe. The following can be shown: an obviousness function $\mathcal{O}$ is globally Boolean iff every agent with $\mathcal{O}$ has a *globally sententially coherent* access table (Corollary 1, Appendix C).

Definition: A probability function P is *globally sententially coherent* iff it satisfies the following conditions for any sentences A, B, X:

S0 $P[\top] = 1$,

S1 $P[\neg A] = 1 - P[A]$,

S2 $P[A \vee B] = P[A] + P[B] - P[A\&B]$,

S3 $P[X] \leq P[B]$ if $X \models B$.

Definition: An access table $\{P_c\}$ is *globally sententially coherent* iff P_c is globally sententially coherent for any choice condition c.

Notation: For any sentence A, $[A]$ is the proposition that A is true. Quotation marks and double brackets are omitted to improve readability. For instance, we write "P_c[It is sunny]" instead of "P_c(["It is sunny"])".

Global sentential coherence amounts to logical omniscience. In the present setup, it is equivalent to standard sentence-based axiomatizations of probability theory (Proposition 1, Appendix C). Agents who are sententially coherent immediately recognize any Boolean connection given any prompt.

Now return to our original question: How severe a constraint is the requirement that an agent's obviousness function is *locally* Boolean? The answer is: moderately severe. The following can be shown: an obviousness function $\mathcal{O}$ is locally Boolean iff every agent with $\mathcal{O}$ has a *locally sententially coherent* access table (Proposition 2, Appendix C).[19,20]

> *Definition:* For c a choice condition, a probability function P_c is *locally sententially coherent* iff it satisfies S0–S3 for all sentences A, B that are salient at c, and all X equal to $\ulcorner A\&B\urcorner$ or $\ulcorner B\&A\urcorner$.
>
> *Definition:* An access table $\{P_c\}$ is *locally sententially coherent* iff P_c is locally sententially coherent for any choice condition c.

In contrast to global sentential coherence, local sentential coherence does not entail logical omniscience. Instead it amounts to restricted logical competence. Agents who are locally sententially coherent respect *simple* Boolean connections when given the right prompt.

[19]The notion of local sentential coherence was inspired by the "locally coherent views" characterized in Gaifman (2004).

[20]Since local sentential coherence is a fairly demanding requirement, the above result shows that locally Booleanness is, too. We learned this from Sinan Dogramaci, who illustrated it with the following beautiful example. Suppose that an agent is playing poker and realizes that a card is to be dealt from a shuffled standard deck. In the agent's present choice condition (c), the only salient sentences are A ("The card is an ace") and H ("The card is a heart"). If the agent's access table is locally sententially coherent, then S2 guarantees that $P_c[A \vee H] = P_c[A] + P_c[H] - P_c[A\&H]$. But this is a fairly demanding requirement. Indeed, in many evaluative contexts we'd count it as rationally permissible for an agent to have $P_c[A] = 4/52$, $P_c[H] = 13/52$, and $P_c[A\&H] = 1/52$ even if $P_c[A \vee H]$ differs from $16/52$.

This example displays a limitation of the above toy model. A more realistic implementation of the proposal in section 4 would employ a notion of obviousness that better matched the cognitive capabilities of the agent in question. In the case of ordinary humans, we expect the details to be too messy and idiosyncratic to admit of a concise mathematical characterization. We also expect that no single characterization of obviousness will serve all theoretical purposes: different standards of rationality will be appropriate in different discussions.

7 *Logical thought is conditionalization on logical information*

The paper so far has aimed to develop a decision theory suitable for logically limited agents. Access tables represent such agents' credences, FRAGMENTED CHOICE RULE gives a criterion for rational action, and CONNECTIVES requires such agents to recognize some (but not all) logical connections. But logically limited agents also *think*. Our next and final goal is to sketch a simple Bayesian model of logical thought.

Start with an example: an agent comes to recognize the truth of a particular tautology by thought alone. How might we model such a cognitive accomplishment using the above framework?

There are actually two questions here: First, how can a particular train of thought be modeled as a sequence of changes to an access table? Second, what constrains *which* train of thought a rational subject engages in? Here is a proof-of-concept proposal to address the first question:

FRAGMENTED INFERENCE Each agent starts out with a given access table and an obviousness function, $\mathcal{O}$. Whenever she encounters a new choice condition c, logical information $\mathcal{O}_c$ becomes broadly accessible: each probability function in her access table conditionalizes on $\mathcal{O}_c$.[21]

When applied to the case of logical inference, one might apply this idea by saying that each step in a chain of thought renders a particular set of sentences salient. Attending to those sentences makes a batch of logical information available to the subject going forward. Successive batches of information combine to produce increasingly powerful logical insights as the chain of thought continues.

For example, suppose that Moriarty has hatched a plan to take advantage of Watson's rather ordinary intellect. He has arranged for one of his

[21] In other words: for every choice condition d, $P_d(\cdot)$ is replaced by $P_d(\cdot|\mathcal{O}_c)$. It is worth emphasizing that the present proposal is no more than a toy model. For example, to avoid conditionalizing on propositions with probability 0 we assume that for any choice conditions c and d, $P_d(\mathcal{O}_c) > 0$. We also restrict attention to the case of an agent with a perfect memory since the present model contains no mechanism for information loss. (This restriction is for simplicity—the model is congenial to relaxing the perfect memory assumption. We explore this in separate work.) In addition, a more general model would probably need to say that not all rows of an access table conditionalize on new logical information.

associates to offer Watson a bet on a fiendishly complicated sentence:

$$((S\&\neg W) \vee S)\&(\neg(W\&S)\&((S\&W) \vee W)).$$

Watson is informed that the bet would cost him just £5, and would pay him £100 if the complicated sentence is true. In fact, the sentence is a contradiction, but it would be too demanding to require Watson to see this immediately.

Watson must decide whether to take the bet. Initially Watson thinks that the sentence might well be true. His access table at each row relevant to the current decision assigns a not-very-low probability to the truth of the sentence. That is compatible with Watson's satisfying CONNECTIVES, since CONNECTIVES only enforces highly limited local constraints at each choice condition. (We present a model to illustrate this in Appendix D.)

Moriarty hopes that Watson will make a snap decision and immediately pay for the (losing) bet. But Watson instead decides to mull things over. His first step is to analyze the leftmost conjunction in Moriarty's sentence. In the model, this is captured by having Watson start out at a choice condition c_1 at which "S" and "$\neg W$" are salient. According to FRAGMENTED INFERENCE, each row of his access table then conditionalizes on $\mathcal{O}_{c_1}$, which entails: "$S\&\neg W$" is true exactly when "S" and "$\neg W$" are both true.

Now that Watson has a grip on "$S\&\neg W$", he attempts to analyze "$(S\&\neg W) \vee S$". In the model, this is captured by having Watson turn to a choice condition c_2 at which "$S\&\neg W$" and "S" are salient. After further conditionalizing on $\mathcal{O}_{c_2}$, each row of Watson's access table now also includes the following information: "$(S\&\neg W) \vee S$" is true exactly when "S" is.[22]

Watson continues in this way, successively attending to various subsentences. At each step, the rows of his access table conditionalize on the bit of information made available at that step. The cumulative result is that each row conditionalizes on a sequence $\mathcal{O}_{c_1}, \mathcal{O}_{c_2}, \ldots, \mathcal{O}_{c_n}$ of batches of information whose conjunction entails that Moriarty's sentence is false. So at the end of

[22]By saying that a row "includes" some information we mean that the probability function associate with that row assigns probability 1 to the set of worlds at which that information obtains.

the process, each row of Watson's access table assigns probability 0 to the truth of Moriarty's sentence.

Watson declines the bet.

8 *Rational thought is a series of utility-maximizing choices of what to think about next*

We are partway through sketching a Bayesian model of logical thought. As noted above, giving such a model required answering two questions: First, how can a particular train of thought be modeled as a sequence of changes to an access table? Second, what constrains *which* train of thought a rational subject engages in? We addressed the first question in the preceding section. Now turn to the second.

Consider Watson's state of mind right after he is offered Moriarty's bet. It might seem that Watson's options are just to accept the bet or to reject it. But Watson has an additional option: to mull things over. And we can apply the rule of expected utility maximization to evaluate that option's choiceworthiness.[23]

That observation leaves it open what exactly Watson will or should think about. But we can go further. We can more finely individuate Watson's thinking-options. We can say that for certain parts of Moriarity's sentence, Watson has the option of attending to just those parts.[24] For example, here is one of Watson's initial options: attend to just "S" and "$\neg W$". That is the option he in fact chooses. Was that choice—the choice to first attend to *those* subsentences—rational? No additional apparatus is needed to answer that question. We can simply treat Watson's choice of what to think about as a choice like any other—rational if and only if it satisfies FRAGMENTED CHOICE RULE (relative to the choice condition Watson is in before he makes

[23]Compare to Hoek (2019, 130–132), I. J. Good (1968, 93–94), Savage (1967, 308), Pettigrew (2020, §6.2).

[24]Compare: "some actions might be totally mental or computational. For example, some actions might control what an agent chooses to think about, or where it focuses its attention." (Sutton and Barto 2018, 55) For work on "rational meta-reasoning", see for example Matheson (1968), Hay et al. (2012), Lieder and Griffiths (2019), Russell and Wefald (1991), Griffiths et al. (2019).

the choice).

What sort of considerations influence the rationality of thinking a particular thought? Here is an example to convey the flavor of how the model operates.

Suppose that Watson is midway through his train of thought—he has so far attended to only a few subsentences of Moriarity's complex sentence. He now has many options. For example, he could focus his attention on a particular "fresh" formula of the sentence—a formula he had not previously attended to. Alternatively, he could attend to a formula he attended to a few moments ago. He could even focus on a sweet childhood memory. Which option maximizes Watson's expected utility? As in any decision situation, the answer depends on Watson's probability function relative to his current choice condition. Suppose that this function takes it to be highly likely that attending to the fresh formula will bring new logical insights but focusing on an old formula or the sweet memory will not. Then, assuming that Watson sufficiently values money, the option that maximizes his expected utility is to attend to a fresh formula. If Watson continues in this way to make utility-maximizing choices—including choices of what to think about—he will eventually figure out that the complicated sentence is contradictory and foil Moriarty's plot.

Appendices

A Decision theory with incoherent credence functions

In §2 we considered the suggestion that incoherent credences—credences that violate the probability axioms—might deliver a decision theory for logically ignorant subjects. We breezily hinted at how to construct such a theory: "Just remove the assumption of logical omniscience from decision theory".

In fact, matters are not quite so straightforward. For once the coherence conditions are dropped, there are certain theoretical choices that need to be made, and it is not obvious whether they can be made in a principled way. Let us explain.

Assume that credences are assigned to propositions. When propositions are modelled as sets of possible worlds, one immediately gets the result that necessarily equivalent propositions get the same credence. So theorists interested in incoherent credences have often chosen to work with a more fine-grained conception of proposition. There are many different ways of spelling out the details (De Bona and Staffel 2017, Gaifman 1988, Garber 1983, Hacking 1967, Zynda 1996, 196–197), but for concreteness we consider below an approach based on distributions of truth values (Williams 2016). Similar points would apply to other fine-grained conceptions of proposition.

Let a *distribution* be an assignment of truth-values to a finite set of sentences and let $\mathcal{W}$ be the set of all such distributions.[25] Assume, further, that each subject has a credence function r that is a probability function over that set of worlds and a value function v that assigns a real number to each world. For any sets X, Y of worlds such that $r(Y) > 0$, we follow standard practice and let $r(X|Y) = r(XY)/r(X)$ be the conditional probability of X given Y and let $V(Y) = \Sigma_{w \in \mathcal{W}} v(w) r(w|Y)$ be the expected value of Y.

Credences and values are so far defined for propositions understood

[25]Note that many of the members of $\mathcal{W}$ might be thought of as *impossible worlds*. Elliott (2019) convincingly argues that theories based on probability distributions over spaces that include such worlds are under strong pressure to say that most propositions are inexpressible.

as sets of distributions. But they can be defined for sentences as well. For any sentence A, let $[A]$ be the proposition that A is true. Then for any sentences A, B we can write $r[A]$ for $r([A])$, $r[A|B]$ for $r([A]|[B])$, and $V[A]$ for $V([A])$.[26]

Can a decision theory be constructed that assumes that agents have incoherent credence functions of this kind? Anyone constructing such a view faces two important decision points.

First, suppose that an agent has an incoherent credence function r. What choice rule determines which options are rational for that agent to select? A natural answer is: the options that maximize expected values relative to r (or some causal quantity computed relative to r—but set such complications aside here). It is natural in this framework to associate each option with a sentence. But care must be taken in deciding *which* sentence to select. That is because when the subject is in a choice situation, there exist many sentences that might describe each of her options.

Suppose, for example, that the subject must press a piano key with her left hand and that using the ring finger to do so is one of her options. The theorist might choose to represent that option using either of the following sentences:

- "use second finger from left"
- "use fourth finger from right"

Which, if any, of these sentences should she use to calculate expected utility? And on what grounds? Note that the difference might matter a great deal, since there is no guarantee that

$$V[\text{"use second finger from left"}] = V[\text{"use fourth finger from right"}].$$

Second, a proponent of the view we are considering also needs a way of accounting for rational credal updates. Consider a subject with credence function r and let r_E be the subject's updated credence function after receiving some new evidence. From a Bayesian perspective, we would like it to

[26] Here we follow Williams (2018, 2016).

be the case that $r_E[A]$ is equal to $r[A|E]$, where E is a sentence capturing the new evidence. But care must be taken in deciding *which* sentence to select. That is because there exist many sentences that might capture a given batch of evidence.

Suppose, for example, that an ordinary subject who knows nothing about the weather walks outside on a sunny and windy day. The theorist might represent the subject's weather-related evidence using either "sunny and windy" or "windy and sunny" (or some more complicated variant).

But which sentence? And on what grounds? Again, the difference might matter a great deal, since there is no guarantee that

$$r[\,\cdot\,|\text{“sunny and windy”}] = r[\,\cdot\,|\text{“windy and sunny”}].$$

We do not claim or even hint that these questions are unanswerable, but only insist that a decision theory based on incoherent credence functions must answer them somehow.

B Access tables and propositional attitudes

How should one understand the relationship between an agent's propositional attitudes and her access table? The following answer is tempting:[27]

ROW CONFIDENCE An agent is confident in a claim if and only if some row of her access table assigns high probability to that claim.

Unfortunately, this straightforward answer leads to the wrong results. For example, it entails that a puzzlist with the access table of section 3 is confident not just in the claim that *dreamt* is a word of English spelled D-R-E-A-M-T, but also in the negation of that claim.

We would also get the wrong results if we were to substitute "every" for "some" in ROW CONFIDENCE. For example, we would be left with

[27]Compare: Lewis (1996) assumes that "S knows that P iff any one of S's compartments knows that P", and Stalnaker (1984, 83) relies on the premise that "what it means to say that an agent believes that P at a certain time is that some one of the belief states the agent is in at that time entails that P" to show that "[i]t is compatible with the pragmatic account that the rational dispositions that a person has at one time should arise from several different belief states".

the conclusion that the puzzlist fails to be confident that *dreamt* is a word of English spelled D-R-E-A-M-T. But intuitively speaking, the puzzlist *is* confident, it's just that she is unable to bring the relevant information to mind as she works on the puzzle.

In the light of this, how should we understand the relationship between an agent's propositional attitudes and her access table? The answer is that access tables should not be thought of as directly corresponding to propositional attitudes. But this does not entail that there is no connection between them. For access tables and propositional attitudes have something important in common: they can both be used to predict a subject's dispositions. And this gives us a useful heuristic connecting the two:

HEURISTIC Let A be a family of propositional attitudes (a family of beliefs, desires, fears, suspicions, etc.) rich enough to make definite predictions about how a rational subject would be disposed to behave under various circumstances. It is appropriate to ascribe A to a fragmented subject if and only if the dispositions predicted by A are sufficiently similar to the dispositions predicted by the subject's access table (along with a suitable value function, via FRAGMENTED CHOICE RULE).

Here is an example. Suppose a family propositional attitudes A includes the belief that there is beer in the fridge. What sorts of dispositions will A predict? Assuming there is nothing extraordinary about A—assuming it consists of attitudes that might be used to describe an ordinary subject in an ordinary situation—the relevant dispositions might include:

> the disposition to say, in appropriate circumstances, sentences like 'There's beer in my fridge'; the disposition to look in the fridge if one wants a beer; a readiness to offer beer to a thirsty guest; the disposition to utter silently to oneself, in appropriate contexts, 'There's beer in my fridge'; an aptness to feel surprise should one go to the fridge and find no beer; the disposition to draw conclusions entailed by the proposition that there is beer in the fridge e.g., that there is something in the fridge, that there is beer in the house; and so forth. (Schwitzgebel 2002, 251)

So, according to HEURISTIC, it is appropriate to ascribe A to a fragmented subject only if her access table predicts dispositions of this kind.

Note that Schwitzgebel's list includes *linguistic* dispositions: dispositions whose manifestation involves the deployment of linguistic information. And when we turn our attention from the belief that there is beer in the fridge to logical and mathematical beliefs, linguistic dispositions can be expected to play an especially prominent role. Suppose, for example, that A includes Frege's ill-fated belief that every concept has an extension. What sorts of dispositions should one expect A to predict? It is easy to think of dispositions that make essential use of linguistic information; for instance: the disposition to utter (in appropriate circumstances) a sentence expressing the claim that every set has an extension, and the disposition to make inferences that rely on the truth of such sentences.[28] But it is harder to come up with relevant dispositions that do not rely on linguistic information. And the more complex and abstract the mathematical belief, the harder it will typically be.

So: logical and mathematical beliefs tend to be tightly linked to linguistic dispositions. But note that this does not entail that they have language as their subject matter. Compare: the belief that there is beer in the fridge might be associated with linguistic dispositions even though it does not have language as its subject matter.

With this as our background, let us return to one of the questions in the main text: Didn't Frege have mistaken beliefs about the nature of extensions, not just mistaken beliefs about language? For example, didn't he mistakenly believe that every (Fregean) concept has an extension? Answer: of course he did! But having that belief doesn't require any row of Frege's access table to assign high probability to the (empty) set of worlds in which every concept has an extension. According to HEURISTIC, what is required is that his access table predict the right dispositions. In this case, the relevant dispositions are mainly linguistic, even though the subject matter of the belief in question is not.

[28] Compare to Schwitzgebel (2013, 89).

C Proofs that (local) Booleanism is equivalent to (local) sentential coherence

Let $\mathscr{L}$ be the set of sentences of a propositional calculus closed under &, $\vee$, and $\neg$, and including a zero-place connective $\top$. We use $\models$ to express logical implication for sentences of $\mathscr{L}$.

Let $\mathcal{W}$ be a space of possible worlds. Assume that there is a world for every distribution of truth values to sentences of $\mathscr{L}$. More precisely: A *distribution* on $\mathscr{L}$ is a function from $\mathscr{L}$ to the set of truth values $\{1, 0\}$. For δ a distribution and $\boldsymbol{A}$ a set of sentences of $\mathscr{L}$, let $\delta[\boldsymbol{A}]$ be the set of worlds in $\mathcal{W}$ at which each $A \in \boldsymbol{A}$ gets truth value $\delta(A)$ and assume that $\mathcal{W}$ is such that $\delta[\mathscr{L}] \neq \emptyset$ for every distribution δ. In addition, let $[\boldsymbol{A}]$ be the set of worlds at which each member of $\boldsymbol{A}$ is true (has truth value 1). For $A \in \mathscr{L}$ let $[A] = [\{A\}]$.

Let $\mathscr{F}$ be a sigma-algebra of subsets of $\mathcal{W}$ such that for each finite $\boldsymbol{A} \subseteq \mathscr{L}$, $\delta[\boldsymbol{A}] \in \mathscr{F}$. A probability function on $\langle \mathcal{W}, \mathscr{F} \rangle$ is a function P that assigns a real number to each proposition in $\mathscr{F}$ and satisfies the following conditions for any $X, Y \in \mathscr{F}$:

K1 $P(X) \geq 0$,

K2 $P(\mathcal{W}) = 1$,

K3 $P(X \cup Y) = P(X) + P(Y) \qquad (XY = \emptyset)$.

We indicate set intersection of propositions by concatenation, so that for example "XY" denotes the intersection of propositions X and Y.

Let C be a nonempty fixed set of choice conditions. Assume that at each choice condition c, finitely many sentences of $\mathscr{L}$ are *salient at* c. An *access table* on C is a function $\{P_c\}$ that maps each choice condition $c \in C$ to a probability function P_c on $\langle \mathcal{W}, \mathscr{F} \rangle$.

An *obviousness function* is a map from C to the set of non-empty propositions. An obviousness function $\mathcal{O}$ is *consistent with* an access table iff for all choice conditions c, $P_c(\mathcal{O}_c) = 1$. An obviousness function $\mathcal{O}$ is *locally Boolean* if for all choice conditions c, $\mathcal{O}_c$ entails that the following conditions hold whenever A and B are sentences salient at c:

B0 $\top$ is true,

B1 $\ulcorner \neg A \urcorner$ is true iff A is not true,

B2 $\ulcorner A \vee B \urcorner$ is true iff either A or B is true,

B3 $\ulcorner A\&B \urcorner$ is true iff both A and B are true,

or equivalently:

b0 $[\top]\mathcal{O}_c = \mathcal{O}_c$,

b1 $[\neg A]\mathcal{O}_c = (\mathcal{W} \setminus [A])\mathcal{O}_c$,

b2 $[A \vee B]\mathcal{O}_c = ([A] \cup [B])\mathcal{O}_c$,

b3 $[A\&B]\mathcal{O}_c = [A][B]\mathcal{O}_c$.

A probability function P is *globally sententially coherent* iff it satisfies the following conditions for any sentences A, B, X:

S0 $P[\top] = 1$,

S1 $P[\neg A] = 1 - P[A]$,

S2 $P[A \vee B] = P[A] + P[B] - P[A\&B]$,

S3 $P[X] \leq P[B]$ if $X \models B$.

An access table $\{P_c\}$ is *globally sententially coherent* iff P_c is globally sententially coherent for any choice condition c.

For c a choice condition, a probability function P_c is *locally sententially coherent* iff it satisfies S0–S3 for all sentences A, B *that are salient at* c and all X equal to $\ulcorner A\&B \urcorner$ or $\ulcorner B\&A \urcorner$. An access table $\{P_c\}$ is *locally sententially coherent* iff P_c is locally sententially coherent for any choice condition c.

In §6 we noted that S0–S3 are equivalent to standard sentence-based axiomatizations of probability theory. The following is one such axiomatization:[29]

For any sentences $A, B \in \mathscr{L}$:

[29]For instance, A1–A3 are referred to as "the probability axioms" in Earman (1992, 36).

A1 $q(A) \geq 0$,

A2 $q(A) = 1 \quad$ if $\models A$,

A3 $q(A \vee B) = q(A) + q(B) \quad$ if $\models \neg(A\&B)$.

Proposition 1. *Let P be a probability function on $\langle \mathcal{W}, \mathscr{F} \rangle$ and $q : \mathscr{L} \to [0,1]$ be such that for all $C \in \mathscr{L}$, $q(C) = P[C]$. Then the following are equivalent:*

1. *P is globally sententially coherent.*
2. *q satisfies A1–A3 for all $A, B \in \mathscr{L}$.*

Proof. First we assume (1) and show that (2) holds:

A1: Since P is a probability function, K1 gives us $P[A] \geq 0$. So A1 follows from the fact that $q(A) = P[A]$.

A2: Assume $\models A$. Then $A \models \top$. So by S3 $P[\top] \leq P[A]$. But by S1, $P[\top] = 1$. So $P[A] \geq 1$. Since K1–K3 entail that $P[A] \leq 1$, it follows that $P[A] = 1$. So A2 follows from the fact that $q(A) = P[A]$.

A3: Assume $\models \neg(A\&B)$. Then A2 guarantees that $q(\neg(A\&B)) = 1$ and therefore $P[\neg(A\&B)] = 1$. So, by S1, we have $P[A\&B] = 0$. So S2 entails that $P[A \vee B] = P[A] + P[B]$. So A3 follows from the fact that $q(C) = P[C]$ for arbitrary C.

We now assume (2) and verify that (1) holds. In fact, it will be convenient to verify that P satisfies S0–S4 for any $A, B \in \mathscr{L}$, where

S4 $P[A] = P[B] \quad$ if $A \models\!\!=\!\!\models B$.

The proofs of S0–S4 are as follows:

S0: $q(\top)$ follows from A2 since $\top$ is a tautology. So S0 follows from $q(\top) = P[\top]$.

S1: Since $A \vee \neg A$ is a tautology, A2 gives us $q(A \vee \neg A) = 1$ and therefore $P[A \vee \neg A] = 1$. Since $\neg(A\&\neg A)$ is a tautology, A3 gives us $q(A \vee \neg A) = q(A) + q(\neg A)$ and therefore $P[A \vee \neg A] = P[A] + P[\neg A]$. Putting the two together gives us $1 = P[A] + P[\neg A]$, which entails S1.

S4: Suppose A and B are logically equivalent. Then the following are all tautologies:

$$(a)\ \neg A \vee (A\&B) \qquad (a')\ \neg(\neg A\&(A\&B))$$
$$(b)\ \neg B \vee (A\&B) \qquad (b')\ \neg(\neg B\&(A\&B))$$

By (a), A2 entails $q(\neg A \vee (A\&B)) = 1$ and therefore $P[\neg A \vee (A\&B)] = 1$. By (a'), A3 entails $q(\neg A \vee (A\&B)) = q(\neg A) + q(A\&B)$ and therefore $P[\neg A \vee (A\&B)] = P[\neg A] + P[A\&B]$. Putting the two together gives us $1 = P[\neg A] + P[A\&B]$ and therefore $1 - P[\neg A] = P[A\&B]$, which, by S1, is equivalent to $P[A] = P[A\&B]$. Analogous reasoning based on (b) and (b') gives us $P[B] = P[A\&B]$. So we have $P[A] = P[A\&B] = P[B]$.

S2: S4 entails $P[A \vee B] = P[(A\&B) \vee (\neg A\&B) \vee (A\&\neg B)]$. But any two of $(A\&B)$, $(\neg A\&B)$, and $(A\&\neg B)$ are logically inconsistent. So A3 gives us $q((A\&B) \vee (\neg A\&B) \vee (A\&\neg B)) = q(A\&B) + q(\neg A\&B) + q(A\&\neg B)$ and therefore $P[(A\&B) \vee (\neg A\&B) \vee (A\&\neg B)] = P[A\&B] + P[\neg A\&B] + P[A\&\neg B]$. Putting the two together gives us $P[A \vee B] = P[A\&B] + P[\neg A\&B] + P[A\&\neg B]$. Adding $P[A\&B]$ to both sides and rearranging the terms gives us: $P[A \vee B] + P[A\&B] = P[A\&B] + P[A\&\neg B] + P[A\&B] + P[\neg A\&B]$.

Now, since $(A\&B)$ is inconsistent with each of $(A\&\neg B)$ and $(\neg A\&B)$, A3 yields: $q((A\&B) \vee (A\&\neg B)) + q((A\&B) \vee (\neg A\&B)) = q(A\&B) + q(A\&\neg B) + q(A\&B) + q(\neg A\&B)$, and therefore $P[(A\&B) \vee (A\&\neg B)] + P[(A\&B) \vee (\neg A\&B)] = P[A\&B] + P[A\&\neg B] + P[A\&B] + P[\neg A\&B]$. So bringing in the result of the previous paragraph gives us $P[A \vee B] + P[A\&B] = P[(A\&B) \vee (A\&\neg B)] + P[A\&B \vee (\neg A\&B)]$. By S4, this gives us: $P[A \vee B] + P[A\&B] = P[A] + P[B]$. Rearranging terms we have $P[A \vee B] = P[A] + P[B] - P[A\&B]$, which is S2.

S3: S4 delivers $P[B] = P[(B\&X) \vee (B\&\neg X)]$. And since S0, S1 and S4 together give us $P[(B\&X)\&(B\&\neg X)] = 0$, S2 delivers $P[(B\&X) \vee (B\&\neg X)] = P[B\&X] + [B\&\neg X]$. Putting the two together gives us $P[B] = P[B\&X] + P[B\&\neg X]$. But now suppose that $X \models B$. Then X and $B\&X$ are logically equivalent. So S4 allows us to conclude

$P[B] = P[X] + P[B\&\neg X]$. But, by K1, we have $P[B\&\neg X] \geq 0$. So we may conclude $P[X] \leq P[B]$, which is S3.

□

Proposition 2. *An obviousness function $\mathcal{O}$ is locally Boolean iff every access table consistent with $\mathcal{O}$ is locally sententially coherent.*

Proof. Left-to-right direction. Assume that $\mathcal{O}$ is locally Boolean and let $\{P_c\}$ be any access table consistent with $\mathcal{O}$. We show that $\{P_c\}$ is locally sententially coherent by arguing, for an arbitrary $c \in C$, that P_c satisfies S0–S3 with respect to any A, B that are salient at c and all X equal to $\ulcorner A\&B \urcorner$ or $\ulcorner B\&A \urcorner$. P_c is automatically locally sententially coherent if no sentences are salient at c, so we assume below that some sentence A is salient at c.

S0: By b0, $[\top]\mathcal{O}_c = \mathcal{O}_c$. But since $\mathcal{O}$ is consistent with $\{P_c\}$, we have $P_c(\mathcal{O}_c) = 1$ and therefore $P_c([\top]\mathcal{O}_c) = 1$. So, again using the fact that $P_c(\mathcal{O}_c) = 1$, we have $P_c[\top] = 1$.

S1: By b1, $[\neg A]\mathcal{O}_c = (\mathcal{W} \setminus [A])\mathcal{O}_c = \mathcal{W}\mathcal{O}_c \setminus [A]\mathcal{O}_c$. Also, $\mathcal{O}_c = [A]\mathcal{O}_c \cup [\neg A]\mathcal{O}_c$. So we have:

$$
\begin{aligned}
1 &= P_c(\mathcal{O}_c) && (P_c \text{ is consistent with } \mathcal{O}) \\
&= P_c([A]\mathcal{O}_c \cup [\neg A]\mathcal{O}_c)) && (\mathcal{O}_c = [A]\mathcal{O}_c \cup [\neg A]\mathcal{O}_c) \\
&= P_c([A]\mathcal{O}_c) + P_c([\neg A]\mathcal{O}_c) - P_c([A][\neg A]\mathcal{O}_c) \\
& && \text{(Theorem of the probability calculus)} \\
&= P_c([A]\mathcal{O}_c) + P_c([\neg A]\mathcal{O}_c) && ([A] \cap [\neg A] = \emptyset) \\
&= P_c[A] + P_c[\neg A], && (P_c(\mathcal{O}_c) = 1.)
\end{aligned}
$$

and hence $P_c[\neg A] = 1 - P_c[A]$.

S2: By b2, $[A \vee B]\mathcal{O}_c = ([A] \cup [B])\mathcal{O}_c$ and by b3 $[A\&B]\mathcal{O}_c = [A][B]\mathcal{O}_c$. So

we have

$$
\begin{aligned}
P_c([A \vee B]\mathcal{O}_c) &= P_c(([A] \cup [B])\mathcal{O}_c) && \text{(b2)} \\
&= P_c([A]\mathcal{O}_c \cup [B]\mathcal{O}_c)) \\
&\qquad\qquad (([A] \cup [B])\mathcal{O}_c = ([A]\mathcal{O}_c \cup [B]\mathcal{O}_c)) \\
&= P_c([A]\mathcal{O}_c) + P_c([B]\mathcal{O}_c) - P_c([A][B]\mathcal{O}_c) \\
&\qquad\qquad \text{(Theorem of the probability calculus)} \\
&= P_c([A]\mathcal{O}_c) + P_c([B]\mathcal{O}_c) - P_c([A\&B]\mathcal{O}_c) && \text{(b3)} \\
&= P_c[A] + P_c[B] - P_c[A\&B] && (P_c(\mathcal{O}_c) = 1.)
\end{aligned}
$$

S3: By b3, $[A\&B]\mathcal{O}_c = [A][B]\mathcal{O}_c \subseteq [A]\mathcal{O}_c$. So $P_c([A\&B]\mathcal{O}_c) \leq P_c([A]\mathcal{O}_c)$. Since $P_c(\mathcal{O}_c) = 1$, this entails $P_c[A\&B] \leq P_c[A]$. Similar reasoning shows that $P_c[A\&B] \leq P_c[B]$.

Right-to-left direction: We must show that if every access table consistent with an obviousness function $\mathcal{O}$ is locally sententially coherent, then $\mathcal{O}$ is locally Boolean.

Assume that every access table consistent with $\mathcal{O}$ is locally sententially coherent and suppose for contradiction that $\mathcal{O}$ is not locally Boolean. Then there exists a choice condition c^*, sentences A, B salient at c^*, and X equal to $\ulcorner A\&B \urcorner$ or $\ulcorner B\&A \urcorner$ such that at some world w^* in $\mathcal{O}_c$, at least one of the conditions for local Booleanness fails. Let I_{w^*} be the unique probability function on $\langle \mathcal{W}, \mathscr{F} \rangle$ such that $I_{w^*}(\{w^*\}) = 1$.

The definition of obviousness function guarantees that $\mathcal{O}_c$ is a nonempty proposition for every choice condition c. So there must be some access table $\{P_c^0\}$ that is consistent with $\mathcal{O}$. Define an access table $\{P_c\}$ by stipulating that for any $c \in C$ and any proposition X:

$$
P_c(X) = \begin{cases} .01P_c^0(X) + .99I_{w^*}(X) & \text{if } c = c^*, \\ P_c^0(X) & \text{otherwise.} \end{cases}
$$

By construction $\{P_c\}$ is an access table such that $P_{c^*}(\{w^*\}) > .9$. Also by construction $\{P_c\}$ is consistent with $\mathcal{O}$ and so is locally sententially coherent.

We know that w^* fails to satisfy one of the Boolean connections in B0–B3 above for A, B salient at c^* and X equal to $\ulcorner A\&B \urcorner$ or $\ulcorner B\&A \urcorner$. But which of B0–B3 fails? We verify that none is possible, using "p" to abbreviate "P_{c^*}" and letting v be the unique distribution that obtains at w^*.

B0: Suppose v fails to satisfy B0. In other words: $v(\top) = 0$. Since $p(\{w^*\}) > 0.9$, it follows that $p[\top] < 0.1$, which contradicts S0.

B1: Suppose v fails to satisfy B1 for A salient at c^*. In other words: $v(\neg A) \neq 1 - v(A)$. So $v(A)$ and $v(\neg A)$ are either both 0 or both 1.

Suppose, first, that $v(A)$ and $v(\neg A)$ are both 1. Then, since $p(\{w^*\}) > 0.9$, we have $p[A] > 0.9$ and $p[\neg A] > 0.9$, which contradicts S1 and therefore our assumption that the agent's access table is locally sententially coherent.

Similarly, suppose that $v(A)$ and $v(\neg A)$ are both 0. Then, since $p(\{w^*\}) > 0.9$, we have $p[A] < 0.1$ and $p[\neg A] < 0.1$, again contradicting S1.

B2: Suppose v fails to satisfy B2 for A, B salient at c. In other words: we have $v(A \vee B) \neq \max(v(A), v(B))$. So either $v(A \vee B) = 1$ and $v(A) = 0 = v(B)$, or $v(A \vee B) = 0$ and at least one of $v(A)$ and $v(B)$ is 1.

Suppose, first, that $v(A \vee B) = 1$ and $v(A) = 0 = v(B)$. Then, since $p(\{w^*\}) > 0.9$, we have $p[A \vee B] > 0.9$, $p[A] < 0.1$, and $p[B] < 0.1$. So, by $S2$, we have two quantities smaller than 0.1 (i.e. $p[A]$ and $p[B]$) adding up to more than 0.9, which is impossible.

Now suppose that $v(A \vee B) = 0$ and at least one of $v(A)$ and $v(B)$ is 1. Then, since $p(\{w^*\}) > 0.9$, we have $p[A \vee B] < 0.1$, and at least one of $p[A]$ and $p[B]$ greater than 0.9. In fact, at most one of $p[A]$ and $p[B]$ can be greater than 0.9. For suppose otherwise. By S2, $p[A \vee B] + p[A\&B] = p[A] + p[B]$. But because $p[A \vee B] < 0.1$, the left hand side of the identity must be smaller than 1.1. This contradicts the assumption that $p[A]$ and $p[B]$ are both greater than 0.9, which entails that the right hand side of the identity is greater than 1.8.

So we know that $p[A \vee B] < 0.1$ and that one of $p[A]$ and $p[B]$ is greater than 0.9 and the other is smaller than 0.1. Given that $p[A \vee B] < 0.1$, it follows that the only way for S2 to be satisfied is for $p[A\&B] > 0.1$. This leads to contradiction because the salience of A and B delivers the following instances of S3: $p[A\&B] \leq p[A]$ and $p[A\&B] \leq p[B]$.

B3: Suppose v fails to satisfy B3 for A, B salient at c. So either $v(A\&B) = 0$ and $v(A) = 1 = v(B)$, or $v(A\&B) = 1$ and at least one of $v(A)$ and $v(B)$ is 0.

Suppose, first, that $v(A\&B) = 0$ and $v(A) = 1 = v(B)$. Then, since $p(\{w^*\}) > 0.9$, we have $p[A\&B] < 0.1$, $p[A] > 0.9$, and $p[B] > 0.9$. So $S2$ entails that two quantities greater than 0.9 (i.e. $p[A]$ and $p[B]$) add up to a number smaller than 1.1 (i.e. the sum of $p[A\&B]$ and $p[A \vee B]$), which is impossible.

Now suppose that $v(A\&B) = 1$ and at least one of $v(A)$ and $v(B)$ is 0. Then, since $p(\{w^*\}) > 0.9$, we have $p[A\&B] > 0.9$ and we have that at least one of $p[A]$ and $p[B]$ smaller than 0.1. This leads to contradiction because the salience of A and B delivers the following instances of S3: $p[A\&B] \leq p[A]$ and $p[A\&B] \leq p[B]$.

□

Corollary 1. *An obviousness function $\mathcal{O}$ is globally Boolean iff every agent with $\mathcal{O}$ has a globally sententially coherent access table.*

Proof. The right-to-left direction is analogous to the right-to-left direction of Proposition 2.

For the left-to-right direction, assume $\mathcal{O}$ is globally Boolean and let $\{P_c\}$ be globally sententially coherent. A proof analogous to the left-to-right direction of Proposition 2 shows that P_c satisfies S0–S2 for any sentences A, B. So it suffices to verify that condition S3 holds for any sentences X and B such that $X \models B$.

Suppose X and B are such that $X \models B$. Then the fact that $\mathcal{O}$ is globally Boolean gives us $[X]\mathcal{O}_c \subseteq [B]\mathcal{O}_c$ and therefore $P_c([X]\mathcal{O}_c) \leq P_c([B]\mathcal{O}_c)$. But since $\mathcal{O}$ is consistent with $\{P_c\}$, we have $P_c(\mathcal{O}_c) = 1$ and therefore

$P_c([X]\mathcal{O}_c) = P_c([X])$ and $P_c([B]\mathcal{O}_c) = P_c([B])$. Putting all of this together gives us $P_c[X] \leq P_c[B]$, which is what we wanted.

□

D *A simple model in which* CONNECTIVES *is nontrivially satisfied without logical omnisicence*

Here is a simple model in which CONNECTIVES is compatible with a failure of logical omnsicience with respect to Moriarty's complicated sentence. The setup uses the framework and notation introduced at the beginning of Appendix C. Let P be a countably additive probability measure over $\langle \mathcal{W}, \mathscr{F} \rangle$ that treats the sentences in $\mathscr{L}$ as if their truth values were determined by independent tosses of a fair coin. More precisely, for any distribution δ on $\mathscr{L}$ and any finite set $\boldsymbol{A} \subset \mathscr{L}$ of sentences, assume that $P(\delta[\boldsymbol{A}]) = 2^{-\#\boldsymbol{A}}$, where $\#\boldsymbol{A}$ is the number of members of $\boldsymbol{A}$. By Carathéodory's Extension Theorem there exists such a P.

Let $\mathcal{O}$ be the weakest locally Boolean obviousness function (i.e. the obviousness function such that for each choice condition c, $\mathcal{O}_c$ is the set of worlds at which B0-B3 hold for any A and B salient at c). To enable us to use P to define an access table, we first check that for any c, $P(\mathcal{O}_c) > 0$. Take any c and recall that the set $\boldsymbol{A}$ of sentences salient at c is assumed to be finite. Let δ^* be a distribution on $\mathscr{L}$ under which the Boolean connectives behave standardly with respect to *all* sentences, so that $\delta^*[\boldsymbol{A}] \subseteq \mathcal{O}_c$. By the definition of P, $P(\delta^*[\boldsymbol{A}]) = 2^{-\#\boldsymbol{A}} > 0$. So $P(\mathcal{O}_c) > 0$. We may now define an access table $\{P_c\}$ by stipulating that for each choice condition c: $P_c(\cdot) = P(\cdot|\mathcal{O}_c)$.

Suppose that $\{P_c\}$ is Watson's access table when he first hears about Moriarty's offer of the bet on the fiendish sentence D:

$$((S\&\neg W) \vee S)\&(\neg(W\&S)\&((S\&W) \vee W)).$$

Initially Watson is in a choice condition c in which the set $\boldsymbol{A}$ of sentences salient to him is small, and in particular does not include all of the subsentences of D. If so, then P_c assigns a not-very-small probability to the fiendish sentence's being true. This can be verified by showing that there is

a distribution γ such that $\gamma(D) = 1$ and $\gamma(\boldsymbol{A}) \subseteq \mathcal{O}_c$. For we then have:

$$\begin{aligned}
P_c([D]) &= P([D]|\mathcal{O}_c) && \text{(Def. of } \{P_c\}) \\
&= P([D]\mathcal{O}_c)/P(\mathcal{O}_c) && \text{(Def. of conditional probability)} \\
&\geq P([D]\mathcal{O}_c) && (P(\mathcal{O}_c) \leq 1) \\
&\geq P(\gamma(\boldsymbol{A})\mathcal{O}_c) && ([D] \supseteq \gamma(\boldsymbol{A}), \text{since } \gamma(D) = 1) \\
&= P(\gamma(\boldsymbol{A})) && (\gamma(\boldsymbol{A}) \subseteq \mathcal{O}_c) \\
&= 2^{-\#\boldsymbol{A}}. && \text{(Def. of } P)
\end{aligned}$$

So $P_c([D]) \geq 2^{-\#\boldsymbol{A}}$, which is not very small since by assumption $\#\boldsymbol{A}$ is small.

To show that there exists a suitable γ for each set $\boldsymbol{A}$ of salient sentences that does not include every strict subsentence of D, we can proceed by cases, depending on which subsentence fails to be included in $\boldsymbol{A}$. Since this is routine but lengthy, we give only one example here. Suppose S is not included in $\boldsymbol{A}$. Then we can choose any γ satisfying the following conditions:

1. $\gamma(S) = 0$.
2. $\gamma(W) = 1$.
3. γ assigns truth-values to a complex sentences as a function of its immediate components, in the usual way, *except* that $(S\&\neg W) \vee S$ is assigned truth-value 1 regardless of the truth values of its immediate components.

To verify that $\gamma(D) = 1$, note that $\gamma(\neg(W\&S)\&((S\&W) \vee W)) = 1$. (This is because $\gamma(S) = 0$ and $\gamma(W) = 1$, and because γ assigns a truth value to $\neg(W\&S)\&((S\&W) \vee W)$ using the standard truth-table.) But since we also have $\gamma((S\&\neg W) \vee S) = 1$, condition (3) guarantees that $\gamma(D) = 1$.

To verify $\gamma[\boldsymbol{A}] \subseteq \mathcal{O}_c$, suppose otherwise. Then there is some $w \in \gamma[\boldsymbol{A}]$ that fails to satisfy one of B0-B3 for $A, B \in \boldsymbol{A}$. But, by condition (3), γ assigns truth-values to complex sentences as a function of their immediate components, in the usual way, *except* for the case of $(S\&\neg W) \vee S$. So the only way in which the failure can occur is for w to fail to satisfy $B2$ when $(S\&\neg W)$ and S are members of $\boldsymbol{A}$, which is impossible, since S is not a member of $\boldsymbol{A}$.

Moral: this model shows that Watson can be far from logically omniscient regarding Moriarty's fiendish sentence, even if he nontrivially satisfies CONNECTIVES.

References

Arif Ahmed. *Evidence, Decision and Causality*. Cambridge University Press, 2014.

Jens Christian Bjerring. Impossible worlds and logical omniscience: an impossibility result. *Synthese*, 190(13):2505–2524, September 2013. ISSN 0039-7857, 1573-0964. doi: 10.1007/s11229-011-0038-y.

Cristina Borgoni, Dirk Kindermann, and Andrea Onofri, editors. *The Fragmented Mind*. Oxford University Press, forthcoming.

David Braddon-Mitchell and Frank Jackson. *The Philosophy of Mind and Cognition, Second Edition*. Blackwell, 2007.

Michael Caie. Rational Probabilistic Incoherence. *Philosophical Review*, 122 (4):527–575, January 2013. ISSN 0031-8108, 1558-1470. doi: 10.1215/00318108-2315288.

Christopher Cherniak. Rationality and the structure of human memory. *Synthese*, 57(2):163–186, 1983.

David Christensen. Does Murphy's Law apply in epistemology? Self-doubt and rational ideals. *Oxford Studies in Epistemology*, 2:3–31, 2007.

Mikaël Cozic. Impossible states at work: Logical omniscience and rational choice. In Richard Topol and Bernard Walliser, editors, *Cognitive Economics: New Trends*, volume 280 of *Contributions to Economic Analysis*, pages 47 – 68. Elsevier, 2006. doi: http://dx.doi.org/10.1016/S0573-8555(06)80003-9. URL `http://www.sciencedirect.com/science/article/pii/S0573855506800039`.

Mark Crimmins. Tacitness and virtual beliefs. *Mind and language*, 7(3):240–263, 1992.

Glauber De Bona and Julia Staffel. Graded incoherence for accuracy-firsters. *Philosophy of Science*, 84(2):189–213, April 2017. ISSN 0031-8248, 1539-767X. doi: 10.1086/690715.

Sinan Dogramaci. Rational credence through reasoning. *Philosophers' Imprint*, 18(11):25, 2018a.

Sinan Dogramaci. Solving the problem of logical omniscience. *Philosophical Issues*, 28(1):107–128, October 2018b. ISSN 15336077. doi: 10.1111/phis.12118.

John Earman. *Bayes or bust? A critical examination of Bayesian confirmation theory*. MIT Press, Cambridge, Mass, 1992. ISBN 978-0-262-05046-3.

Kenny Easwaran. Bayesianism II: Applications and criticisms. *Philosophy Compass*, 6(5):321–332, 2011. ISSN 1747-9991. doi: 10.1111/j.1747-9991.2011.00398.x. URL `http://dx.doi.org/10.1111/j.1747-9991.2011.00398.x`.

Andy Egan. Seeing and believing: Perception, belief formation and the divided mind. *Philosophical Studies*, 140(1):47–63, July 2008. ISSN 0031-8116, 1573-0883. doi: 10.1007/s11098-008-9225-1.

Adam Elga and Agustín Rayo. Fragmented decision theory, January 2017.

Adam Elga and Agustín Rayo. Fragmentation and information access. In Cristina Borgoni, Dirk Kindermann, and Andrea Onofri, editors, *The Fragmented Mind*,. Oxford University Press, Forthcoming.

Edward Elliott. Impossible worlds and partial belief. *Synthese*, 196(8):3433–3458, August 2019. ISSN 0039-7857, 1573-0964. doi: 10.1007/s11229-017-1604-8.

Hartry Field. Mental representation. *Erkenntnis*, 13:9–61, July 1978.

Hartry Field. Stalnaker on intentionality: On Robert Stalnaker's "Inquiry". *Pacific Philosophical Quarterly*, 67:98–112, April 1986a.

Hartry Field. Critical notice: Robert Stalnaker, Inquiry. *Philosophy of Science*, 53(3):425–448, 1986b.

Haim Gaifman. A theory of higher-order probabilities. In B. Skyrms and William Harper, editors, *Causation, chance, and credence*. Kluwer, Dordecht, 1988.

Haim Gaifman. Reasoning with limited resources and assigning probabilities to arithmetical statements. *Synthese*, 140(1/2):97–119, May 2004. ISSN 0039-7857. doi: 10.1023/B:SYNT.0000029944.99888.a7.

Daniel Garber. Old evidence and logical omniscience in Bayesian confirmation theory. In John Earman, editor, *Minnesota studies in the philosophy of science*, volume 10, pages 99–131. University of Minnesota Press, 1983.

Allan Gibbard and William Harper. Counterfactuals and two kinds of expected utility. In A. Hooker, J. J. Leach, and E. F. McClennen, editors, *Foundations and Applications of Decision Theory*, pages 125–162. D. Reidel, 1978.

Thomas L Griffiths, Frederick Callaway, Michael B Chang, Erin Grant, Paul M Krueger, and Falk Lieder. Doing more with less: Meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, 29:24–30, October 2019. ISSN 23521546. doi: 10.1016/j.cobeha.2019.01.005.

Ian Hacking. Slightly more realistic personal probability. *Philosophy of Science*, 34(4):311–325, 1967.

Joseph Y. Halpern and Riccardo Pucella. Dealing with logical omniscience. *Artificial Intelligence*, 175(1):220–235, 2011.

Nicholas Hay, Stuart Russell, David Tolpin, and Solomon Eyal Shimony. Selecting computations: Theory and applications. *arXiv:1207.5879 [cs]*, July 2012.

Jaakko Hintikka. Impossible possible worlds vindicated. *Journal of Philosophical Logic*, 4(4):475–484, 1975.

Daniel Hoek. *The Web of Questions*. PhD thesis, New York University, 2019.

David Hume. *A treatise of human nature*. Clarendon Press, Oxford, 1738. Reprint 1966.

I. J. Good. A five-year plan for automatic chess. *Machine intelligence*, 2:89–118, 1968.

Mark Jago. The problem of rational knowledge. *Erkenntnis*, 79:1151–1168, 2014. ISSN 0165-0106.

Richard Jeffrey. *The Logic of Decision*. University of Chicago Press, 1965.

James M. Joyce. *The Foundations of Causal Decision Theory*. Cambridge University Press, 1999.

Mark Kaplan. *Decision Theory as Philosophy*. Cambridge University Press, Cambridge, 1996.

David Lewis. Causal decision theory. *Australasian Journal of Philosophy*, 59 (1):5–30, 1981.

David Lewis. Logic for equivocators. *Nous*, 16:431–441, 1982.

David K. Lewis. Elusive knowledge. *Australasian Journal of Philosophy*, 74(4): 549–567, 1996. doi: 10.1080/00048409612347521.

Falk Lieder and Thomas L. Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, pages 1–85, February 2019. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X1900061X.

Barton L. Lipman. Decision theory without logical omniscience: Toward an axiomatic framework for bounded rationality. *The Review of Economic Studies*, 66(2):339–361, 1999. ISSN 00346527, 1467937X. URL `http://www.jstor.org/stable/2566994`.

James E. Matheson. The economic value of analysis and computation. *IEEE Transactions on Systems Science and Cybernetics*, 4(3):325–332, September 1968. ISSN 2168-2887. doi: 10.1109/TSSC.1968.300126.

Aaron Norby. Against fragmentation. *Thought*, 2(4):30–38, 2014.

Rohit Parikh. Sentences, belief and logical omniscience, or what does deduction tell us? *Review of Symbolic Logic*, 1(4):459–476, 2008.

Richard G. Pettigrew. Logical ignorance and logical learning. *Synthese*, pages 1–30, 2020. doi: 10.1007/s11229-020-02699-9.

Lawrence Powers. Knowledge by deduction. *Philosophical Review*, 87(3): 337–371, 1978.

Jake Quilty-Dunn and Eric Mandelbaum. Against dispositionalism: Belief in cognitive science. *Philosophical Studies*, 175(9):2353–2372, September 2018. ISSN 1573-0883. doi: 10.1007/s11098-017-0962-x.

Bertrand Russell. Letter to Frege. In Jean van Heijenoort, editor, *From Frege to Gödel: A source book in mathematical logic, 1879-1931*. Harvard Univ. Pr, Cambridge, Mass, 4. pr edition, 1902. ISBN 978-0-674-32449-7.

Stuart Russell and Eric Wefald. Principles of metareasoning. *Artificial Intelligence*, 49(1):361–395, May 1991. ISSN 0004-3702. doi: 10.1016/0004-3702(91)90015-C.

Leonard Savage. *The foundations of statistics*. John Wiley and Sons, 1954.

Leonard J. Savage. Difficulties in the theory of personal probability. *Philosophy of Science*, 34(4):305–310, 1967.

Mark J. Schervish, Teddy Seidenfeld, and Joseph Kadane. Measures of incoherence: How not to gamble if you must, with discussion. In J. M. Bernardo, editor, *Bayesian Statistics*, volume 7. Oxford, New York, 2003.

Eric Schwitzgebel. A phenomenal, dispositional account of belief. *Noûs*, 36 (2):249–275, 2002.

Eric Schwitzgebel. A Dispositional Approach to Attitudes: Thinking Outside of the Belief Box. In Nikolaj Nottelmann, editor, *New essays on belief: Constitution, content and structure*, pages 75–99. Palgrave Macmillan UK,

London, 2013. ISBN 978-1-349-43922-5 978-1-137-02652-1. doi: 10.1057/9781137026521_5.

Teddy Seidenfeld, Mark J. Schervish, and Joseph B. Kadane. What kind of uncertainty is that? Using personal probability for expressing one's thinking about logical and mathematical propositions:. *Journal of Philosophy*, 109(8):516–533, 2012. doi: 10.5840/jphil20121098/925.

Mattias Skipper and Jens Christian Bjerring. Bayesianism for non-ideal agents. *Erkenntnis*, January 2020. ISSN 0165-0106, 1572-8420. doi: 10.1007/s10670-019-00186-3.

Declan Smithies. Ideal rationality and logical omniscience. *Synthese*, 192(9): 2769–2793, 2015. doi: 10.1007/s11229-015-0735-z.

Scott Soames. Understanding assertion. In Scott Soames, editor, *Philosophical Essays, Volume 2: The Philosophical Significance of Language*, pages 211–242. Princeton University Press, 2009.

Jeff Speaks. Is mental content prior to linguistic meaning? *Noûs*, 40(3): 428–467, 2006.

Julia Staffel. Measuring the overall incoherence of credence functions. *Synthese*, 192(5):1467–1493, May 2015. ISSN 0039-7857, 1573-0964. doi: 10.1007/s11229-014-0640-x.

Robert Stalnaker. Letter to David Lewis. In William Harper, Robert Stalnaker, and Glenn Pearce, editors, *Ifs: conditionals, belief, decision, chance, and time*, pages 151–152. Reidel, Dordrecht, 1981. Letter dated 1972.

Robert Stalnaker. *Inquiry*. MIT Press, Cambridge, Mass, 1984. ISBN 978-0-262-19233-0.

Robert Stalnaker. Replies to Schiffer and Field. *Pacific Philosophical Quarterly*, 67(2):113, 1986.

Robert Stalnaker. The problem of logical omniscience. *Synthese*, 89(3):425–440, 1991.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass., second edition, 2018.

J. Robert G. Williams. Rational illogicality. *Australasian Journal of Philosophy*, 96(1):127–141, 2018. doi: 10.1080/00048402.2017.1323933.

Robert Williams. Probability and nonclassical logic. In Alan Hajek and Christopher Hitchcock, editors, *The oxford handbook of probability and philosophy*. Oxford university press, 2016.

Timothy Williamson. Absolute provability and safe knowledge of axioms. In Leon Horsten and Philip Welch, editors, *Gödel's Disjunction: The scope and limits of mathematical knowledge*, chapter 10, pages 243–253. Oxford University Press, Oxford and New York, 2016.

Seth Yalcin. *Modality and inquiry*. PhD thesis, MIT, 2008.

Seth Yalcin. Figure and ground in logical space. Posted at `http://escholarship.org/uc/item/11c0x4n5`, on 12–02–2015, 2015.

Seth Yalcin. Belief as question-sensitive. *Philosophy and Phenomenological Research*, 2016. ISSN 1933-1592. doi: 10.1111/phpr.12330. URL `http://dx.doi.org/10.1111/phpr.12330`.

Lyle Zynda. Coherence as an ideal of rationality. *Synthese*, 109(2):175–216, 1996.